Shifted Eisenstein Polynomials, Irreducible Compositions of Polynomials and Group Key Exchanges

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Reto Alexander Schnyder

von

Rapperswil-Jona SG

Promotionskommission

Prof. Dr. Joachim Rosenthal (Vorsitz)
Prof. Dr. Andrew Kresch
Prof. Dr. Elisa Gorla

Zürich, 2017

Contents

Abstract

In my dissertation, I have covered multiple different topics. First, we consider the concept of natural density over the integers, and extend it to holomorphy rings over function fields. This allows us to give a function field analogue of Cesàro's theorem, which gives the "probability" that an m-tuple of random elements of the holomorphy ring is coprime. We also generalize this and consider the density of $k \times m$ matrices over holomorphy rings which can be extended to unimodular $m \times m$ matrices.

In the second part, we determine the natural density of shifted Eisenstein polynomials. This means that we compute the density of integer polynomials $f(x)$ of a fixed degree n for which some shift $f(x + i)$ for an integer i satisfies Eisenstein's irreducibility criterion. We then also compute the density of affine Eisenstein polynomials.

Thirdly, we consider an arbitrary set of monic quadratic polynomials over a finite field and ask ourselves which compositions of copies of them are irreducible. We first give a criterion to decide whether all such compositions are irreducible, and then show that in general, the irreducible compositions have the structure of a regular language.

In the final chapter, we study cryptographic protocols for key exchange in ad-hoc groups. We first translate some protocols from the literature to the more general setting of semigroup actions, and then propose our own variants of these protocols, which aim to have improved security or efficiency. Then, we demonstrate a couple of active attacks on certain such protocols which are in some ways more powerful than man-in-the-middle attacks.

Acknowledgments

I extend my thanks to everyone who has helped me complete this thesis. First and foremost, I am grateful to my advisor, Joachim Rosenthal, for his guidance and support.

I would like to thank my coauthors, particularly Giacomo Micheli, who was always full of ideas for new questions to tackle. I am also grateful to Juan Antonio López-Ramos and the Departemento de Matemáticas in Almería for allowing me to visit for two weeks.

Furthermore, I thank the Institute of Mathematics at the University of Zürich, who have made for a pleasant work environment, as well as my colleagues and officemates. I also thank Armasuisse, who have provided the funding for my dissertation.

Last but not least, I wish to thank my family, who have always supported me.

Chapter 1

Introduction

1.1 Background on Cryptography

Cryptography is the theory of secure communication. This primarily encompasses two goals. The first is *confidentiality*, that is, ensuring that no third party can listen in on a communication against the will of the communication partners. This is done via the process of *encryption*. The second is *authenticity*, which means preventing adversaries from modifying messages in a communication, and from making up messages entirely and attributing them to another party. This is typically done with *digital signatures* or *message authentication codes*.

Cryptographic schemes can broadly be divided into two categories: *Symmetric cryptography* and *public key cryptography*. In symmetric cryptography, the participants in a communication share a common secret key, which is then used to encrypt and authenticate all messages between them. Symmetric ciphers like AES are typically very fast and well-trusted. Their main weakness however is the need for a shared secret between the parties, which needs to be exchanged in some other secure way beforehand. In some situations, this can be very difficult or downright impossible.

In public key cryptography, on the other hand, the parties do not share a common secret. Rather, each party possesses a key pair consisting of a *public* and a *private key*. As the names suggests, the public key is known to the other parties, whereas the private key is known only to its owner. Encryption is then done as follows: The sender of a message encrypts it using the public key of the receiver, and the receiver can decrypt this message using his private key. If the cipher is secure, the encrypted message cannot be read by anyone who does not possess the corresponding private key. Authenticity is achieved via digital signatures: The sender of a message uses his private key to create a signature for his message, and sends it alongside the message. Anyone who has his public key can then verify this signature. If the signature scheme is secure, no one can create a signed message that is accepted by a given public key unless he knows the corresponding private key. Commonly, encryption and authentication are used together on the

same messages. One of the most used public key cryptosystems is RSA, which can be used for both encryption and signing. Other cryptosystems can often only do one or the other, so they need to be used in pairs.

In practice, public key cryptography is usually not used to encrypt and sign longer communications, since it is comparatively slow. Instead, it is used in conjunction with symmetric cryptography: A public key cryptosystem is used to securely exchange a shared secret key between the parties, which is then used with a symmetric cipher to encrypt and authenticate the rest of the communication.

A common way of doing this is with *key exchange protocols*, whose sole purpose is establishing a shared key between multiple parties. The most famous such protocol is the *Diffie-Hellman key exchange* between two parties (commonly called Alice and Bob).

Protocol 1.1 (Diffie-Hellman Key Exchange [DH76])**.** Both parties agree on a (publicly known) cyclic group G of order n generated by g. Then, Alice chooses a random integer $a \in \{0, \dots, n-1\}$, which she keeps secret, and sends $A = g^a$ to Bob. Likewise, Bob chooses a secret $b \in \{0, \dots, n-1\}$ and sends $B = g^b$ to Alice. After this, Alice computes the shared secret B^a, whereas Bob computes A^b. Clearly, both of these are equal to g^{ab}.

The Diffie-Hellman key exchange is secure if and only if the *Diffie-Hellman problem* is hard in the group G, that is, given the elements g, g^a and g^b of G, one cannot realistically compute g^{ab}. An obvious requirement for this is the intractability of the *discrete logarithm problem* (DLP):

Problem 1.2 (Discrete Logarithm Problem)**.** Given a group G and two elements $g, h \in G$ with $h \in \langle g \rangle$, determine an integer a such that $g^a = h$.

Examples for groups where these problems are thought to be difficult in general are the multiplicative group of a finite field and the additive group of points on an elliptic curve.

It should be noted that the Diffie-Hellman protocol is not authenticated, and an active attacker (call her Mallory) who can intercept and insert her own messages into the channel between Alice and Bob can perform a *man-in-the-middle attack*. This works as follows: Mallory chooses her own secret $m \in \{0, \dots, n-1\}$ and public element g^m. When Alice and Bob exchange their public elements g^a and g^b, Mallory replaces both with g^m. Alice then ends up with the secret g^{am} and Bob with g^{bm}, both of which Mallory can compute since she knows g^a, g^b and m. At this point, Mallory shares a secret key with each of Alice and Bob. When Alice now sends a message encrypted with the secret g^{am} to Bob,[1] Mallory intercepts and decrypts it, reencrypts it with g^{bm}, and sends the result to Bob. She proceeds analogously if Bob sends a message to Alice. In this way, Mallory can read and

[1] More precisely, encrypted with a secret key derived deterministically from g^{am}, to which both Alice and Mallory have access.

even modify all communication between Alice and Bob without them noticing the intrusion.

A large part of my dissertation is concerned with key exchange protocols for groups of more than two users. While it is possible to simply use multiple iterations of the Diffie-Hellman key exchange in this larger setting, we are interested in creating more efficient and dynamic protocols.

1.2 Background on Natural Density

A question that occasionally comes up in number theory is the following: Suppose we are given a set of integers $A \subseteq \mathbb{Z}$. What proportion of all integers lie in A? Or written probabilistically, what is the probability that a randomly selected integer lies in A?

For example, the probability that a random integer is even should intuitively be $\frac{1}{2}$. Likewise, a random integer should be a prime with probability 0 by the prime number theorem. How can we formulate such a question rigorously? Clearly, by cardinality, the sets of even or prime integers have the same size as $\mathbb{Z}$ itself, so this is not a useful approach. Furthermore, it is well known that there exists no uniform probability distribution on the set of integers, so a probability theoretical approach is also impossible.

However, we can do the following: If we consider only integers bounded by some constant B, we get a finite set, and we can compute the proportion of them that lie in A: $|A \cap [-B, B[|/2B$. If we now take the limit as B goes to infinity, and it happens to converge, we call the result the *natural density* of A.

We are primarily interested in the density of sets of tuples of integers $A \subseteq \mathbb{Z}^d$. The definition of this is a straightforward generalization.

Definition 1.3. For any subset $A \subseteq \mathbb{Z}^d$, we define

$$\overline{\mathbb{D}}(A) \coloneqq \limsup_{B\to\infty} \frac{\left|A \cap [-B, B[^d\right|}{(2B)^d},$$
$$\underline{\mathbb{D}}(A) \coloneqq \liminf_{B\to\infty} \frac{\left|A \cap [-B, B[^d\right|}{(2B)^d}.$$

If these coincide, we denote their value by $\mathbb{D}(A)$ and call it the *natural density* of A:

$$\mathbb{D}(A) \coloneqq \lim_{B\to\infty} \frac{\left|A \cap [-B, B[^d\right|}{(2B)^d}.$$

Remark 1.4. It is certainly possible that the limit above does not converge. For an example, take the set $A \subseteq \mathbb{Z}$ of integers whose decimal expansion starts with the digit 1. It can be easily seen that $\underline{\mathbb{D}}(A) = 1/9$ and $\overline{\mathbb{D}}(A) = 5/9$. Hence, the set A does not possess a natural density [TI95, p. 261].

As an example, consider the following question, which has first been answered by Cesàro and Sylvester [Ces83; Syl83]: What is the density of the set of coprime pairs of integers, $A \coloneqq \{(a, b) \in \mathbb{Z}^2 \mid \gcd(a, b) = 1\}$? It turns out that this density is exactly $\zeta(2)^{-1} = 6/\pi^2$, where ζ is the Riemann zeta function.

This can be seen informally as follows: For each prime p, the probability that both a and b are multiples of p is p^{-2}. Hence, the probability that p is not a common divisor of a and b is $1 - p^{-2}$. Now, a and b are coprime if and only if this is the case for all primes, so

$$\mathbb{D}(A) = \prod_{p \text{ prime}} (1 - p^{-2}) = \zeta(2)^{-1}.$$

See e.g. [Nym72] for a real proof. This result can easily be extended to coprime m-tuples of integers, where the density ends up being $\zeta(m)^{-1}$.

1.3 Overview of the Dissertation

For my dissertation, I chose to work in the areas of cryptography and number theory. In cryptography, my focus was on public key cryptography. This led to the study of group key exchanges, where multiple users try to share a common key in an ad-hoc setting. My research in algebra and number theory was also largely inspired by concerns from public key cryptography. In particular, multiple of my papers concern irreducible polynomials, which are important for constructing finite fields of non-prime cardinality.

I will now give a short overview of my dissertation, giving a small introduction to each chapter. Each chapter also possesses its own introduction section, where I will go into more detail.

In Chapter 2, the focus is on an extension of the notion of natural density to holomorphy rings in function fields. In particular, we consider an analogue of Cesàro's theorem, which says that that the density of coprime m-tuples of integers is $\zeta(m)^{-1}$, where ζ is the Riemann zeta function [Nym72]. This theorem has already been extended to rings of polynomials over finite fields by [ST07; GY13]. Our goal then is to further generalize this case to not just polynomial rings, but arbitrary holomorphy rings of function fields over finite fields.

Going with the notations and definitions of [Sti09], a *function field* (of one variable) F is defined as the field of rational functions defined on a smooth projective algebraic curve over a field K. Equivalently, it is a finitely generated field extension of K of transcendence degree one. A valuation ring of F is a subring $K \subsetneq \mathcal{O} \subsetneq F$ such that for each element $z \in F$, at least one of z and z^{-1} lies in $\mathcal{O}$. These rings are local, and the maximal ideal of such a valuation ring is called a *place* of the function field F. We call the set of all places $\mathscr{P}$. Such a place P corresponds to a point on the curve, and its valuation ring $\mathcal{O}_P$ consists of all rational functions on the curve that are defined at that point. A *holomorphy ring* is then

given by an arbitrary intersection

$$H = \bigcap_{P \in \mathcal{S}} \mathcal{O}_P,$$

for some subset $\emptyset \neq \mathcal{S} \subsetneq \mathcal{P}$. It consists of all rational functions on the curve that are defined at all points in $\mathcal{S}$. For a concrete example, on the projective line $\mathbb{P}^1$, we have $F = K(x)$, and the holomorphy ring of $\mathcal{S} = \mathcal{P} \setminus \{\infty\}$ is exactly the polynomial ring $K[x]$.

Our task was now to define an appropriate notion of density in holomorphy rings over finite fields, and to find the density of m-tuples over such rings that are coprime. We further generalize this question to the density of $k \times m$-matrices which can be extended to unimodular matrices, and we are able to answer it when $2k - 1 \leq m$. This chapter is based on joint work with Giacomo Micheli that has been published in [MS16a] and [MS16c].

In Chapter 3, we consider a different question relating densities and polynomials: What is the density of integer polynomials of a fixed degree, considered as tuples, for which the Eisenstein criterion [Eis50] applies? The Eisenstein criterion says that for an integer polynomial $f(x) = a_n x^n + \cdots + a_0$, if there is a prime p that divides a_i for $i < n$ but not a_n, and p^2 does not divide a_0, then $f(x)$ is irreducible over the rationals. The density of such polynomials was computed exactly by [Dub03] and [HS13]. In [HS14], the question was extended to polynomials $f(x) \in \mathbb{Z}[x]$ for which some linear shift $f(x + a)$ satisfies the Eisenstein criterion. That paper however only gave very rough bounds on this density. In our work, we were able to give the exact density of such *shifted Eisenstein polynomials*, as well as the more general affine Eisenstein polynomials. For this, we use a local to global principle for densities given in [PS99, Lemma 20]. This chapter is based on the paper [MS16b], which is joint work with Giacomo Micheli.

In Chapter 4, we have studied the irreducibility of compositions of polynomials. A polynomial over a finite field $\mathbb{F}_q$ is called *stable* if composing it with itself any number times results in an irreducible polynomial. These polynomials have been studied by a variety of authors. For example, in [JB12], a criterion is given to determine whether a quadratic polynomial is stable with only a finite amount of computation. In this chapter, we extend this question to sets $\mathcal{S}$ of multiple monic quadratic polynomials: We determine with a finite amount of computation whether all possible compositions of copies of elements of $\mathcal{S}$ are irreducible. In other words, we consider the compositional monoid generated by $\mathcal{S}$ and ask whether it consists only of irreducible polynomials. Even if this isn't the case, we then show how to use the theory of finite state automata to give a precise description of which elements of this monoid are irreducible. We do this by showing that they have the structure of a regular language. This chapter is based on the work I did with Andrea Ferraguti and Giacomo Micheli that is published in [FMS16] and [FMS17].

Finally, Chapter 5 concerns cryptographic group key exchange protocols. Suppose you have a group of users or other entities $\{\mathcal{U}_1, \ldots, \mathcal{U}_n\}$ who have never

communicated before, and they wish to establish a secure communication channel. Such scenarios are becoming increasingly common with the rise of the so called Internet of things, in which a variety of small sensors and devices need to connect and communicate with one another. This may take place on the scale of a household, a city, or even beyond. However, such networks also pose a new challenge to security: For example, an unauthorized party being able to read the communication between such devices may threaten the privacy of a household. If that party can even inject their own messages into the network, direct physical damage can be done: If, for example, the heating in a household is controlled by sensors throughout the apartment which are connected wirelessly to a thermostat, a hostile party could fake sensor data to cause overheating, damaging not only material goods but potentially the health of the inhabitants. This makes it necessary for the communication channels between the nodes to be encrypted and authenticated. Furthermore, these devices often only possess very weak computational power, so it is important that the protocols involved be as efficient as possible.

We only concern ourselves with unauthenticated exchanges here, assuming that it is not feasible for the parties to share previous authentication information. After establishing the initial shared key, it should be possible for users to join and leave the group later on, for which the key needs to be updated. There are a multitude of protocols that aim to achieve this, but the ones we focus on are primarily based on the works of Steiner et al. in [STW96; STW00], which extend the traditional Diffie-Hellman protocol. In this chapter, we first extend these protocols to the setting of general semigroup actions, and then introduce our own variants which aim to improve either security or efficiency of those protocols. Some of those make use of additional linear structure of the semigroup actions. Afterwards, we present two active attacks, one on a protocol of [STW96] and one on the Burmester-Desmedt protocol of [BD94]. These attacks are in some ways more powerful than a basic man-in-the-middle attack.

Our protocols were published in [Lóp+15] and [Lóp+16], and the first attack in [Sch+16]. These papers were written in collaboration with Juan Antonio López-Ramos, Joachim Rosenthal and Davide Schipani. The second attack is joint work with Mohamed Baouch, Juan Antonio López-Ramos and Blas Torrecillas, and is published in [Bao+16].

Chapter 2

Densities over Holomorphy Rings

2.1 Introduction

A classical result by Ernest Cesàro and James Joseph Sylvester [Ces81; Ces83; Ces84; Syl83] in number theory concerns the "probability" that a random pair of integers is coprime. Formally, as we discussed in Section 1.2, the natural density of the set of coprime pairs of integers is $\frac{1}{\zeta(2)}$, where ζ is the Riemann zeta function. Furthermore, this has been generalized to coprime m-tuples, which have density $\frac{1}{\zeta(m)}$ [Nym72]. Similar results also hold in the rings of integers of arbitrary number fields [FM15; Sit10].

Our goal is to extended this result to holomorphy rings of function fields over finite fields. Additionally, we generalize the question and consider the density of $k \times m$-matrices which can be extended to unimodular $m \times m$-matrices. Our results on these two questions were published in [MS16a] and [MS16c] respectively, on which this chapter of my thesis is based.

Let $\mathbb{F}_q$ be a finite field with q elements and let F be an algebraic function field[1] having full constant field $\mathbb{F}_q$. Let $\mathscr{C}$ be the set of places of F and $\mathcal{S} \subsetneq \mathscr{C}$ be a nonempty proper subset. The *holomorphy ring* of $\mathcal{S}$ is

$$H \coloneqq \bigcap_{P \in \mathcal{S}} \mathcal{O}_P,$$

where $\mathcal{O}_P$ is the valuation ring of the place P. As is well known, these rings are integrally closed and all integrally closed subrings of F are of this form. We say that an m-tuple of elements of H is coprime if its components generate the unit ideal in H, in analogy to the case of the ring of integers in [FM15].

In Section 2.2, we provide a short introduction on the basics of algebraic function fields, concentrating on the definitions we require for the rest of the chapter.

[1] In this chapter, we will mostly use the language and notation of [Sti09].

In Section 2.3, we define an appropriate notion of density for subsets of H^m, using Moore-Smith convergence for nets [Kel55, Chapter 2]. Then, in Section 2.4, we study the density of the set of coprime m-tuples in H, considered as a subset of H^m. We are able to show that this density exists and is equal to $\frac{1}{\zeta_H(m)}$, where ζ_H is the zeta function of the holomorphy ring.

In Section 2.4.1, we provide an example in the case of the affine ring of coordinates of an elliptic curve to show a concrete application of the main result. Then, in Section 2.4.2, we consider the special case $F = \mathbb{F}_q(x)$ and $H = \bigcap_{P \neq P_\infty} \mathcal{O}_P = \mathbb{F}_q[x]$, which has been studied for $m = 2$ in [ST07] and more generally in [GY13]. We will explain how to interpret the densities presented in these papers as particular cases of our general framework.

After this, we turn to the density of $k \times m$-matrices that can be extended to unimodular $m \times m$-matrices. This generalizes to an extent the work of [GY13]. We will state our result in Section 2.5. In Section 2.5.1, we will give an upper bound that works for all $k < m$, and in Section 2.5.2, we will prove that bound exact when $2k - 1 \leq m$.

2.2 Algebraic Function Fields

We will now give a short introduction on the basics of algebraic function fields in one variable. See [Sti09] for more details.

Let K be a field. A *function field* (of one variable) over K is a finite algebraic field extension $F \supseteq K(x)$. Here, $K(x)$ is the rational function field over K, which means that x is transcendental over K. The *field of constants* $\tilde{K}$ of F is the algebraic closure of K in F, that is, $\tilde{K} \coloneqq \{z \in F \mid z \text{ is algebraic over } K\}$. We say that F has *full constant field* K if $\tilde{K} = K$.

In this chapter, we will only be considering the case where K is a finite field and F has full constant field K.

A *valuation ring* $\mathcal{O}$ of the function field F over K is a subring $K \subsetneq \mathcal{O} \subsetneq F$ such that for each $z \in F$, either z or z^{-1} lies in $\mathcal{O}$. Valuation rings are local rings, and we call their maximal ideals *places*. There is a one to one correspondence between valuation rings and places: Places are principal ideals, and if the maximal ideal P of $\mathcal{O}_P$ is generated by the element t, we define the *discrete valuation* at P as the map $v_P : F^* \longrightarrow \mathbb{Z}$ where $v_P(z)$ is the unique integer n such that $t^{-n}z \in \mathcal{O}_P^*$. Given just the place P, its valuation ring can now be recovered as $\mathcal{O}_P = \{z \in F \mid v_P(z) \geq 0\}$.

We write $\mathscr{C}$ for the set of all places of F. A *divisor* of F over K is a finite linear combination of places in $\mathscr{C}$ with coefficients in $\mathbb{Z}$. The divisors form an abelian group, which we denote by $\mathrm{Div}(F)$. The degree of a divisor $D = \sum_{P \in \mathscr{C}} n_P P$ is defined as $\deg(D) \coloneqq \sum_{P \in \mathscr{C}} n_P \deg(P)$, where $\deg(P) \coloneqq \dim_K(\mathcal{O}_P/P)$. The *support* of D is $\mathrm{supp}(D) \coloneqq \{P \in \mathscr{C} \mid n_P \neq 0\}$. We also define a partial order on $\mathrm{Div}(F)$ by defining $D \leq E$ whenever $E - D$ has only nonnegative coefficients.

To an element $z \in F^*$, we can now assign the divisor $(z) \coloneqq \sum_{P \in \mathscr{C}} v_P(z)P$. It can be uniquely decomposed as $(z) = (z)_0 - (z)_\infty$, where $(z)_0$ and $(z)_\infty$ have disjoint

support and each has only nonnegative coefficients. These divisors are called the *zero divisor* and the *pole divisor* of z, respectively.

It is important to note the geometric point of view of all this: An algebraic function field F corresponds to the field of rational functions of some smooth projective algebraic curve C over K. The places of F correspond to the points of this curve over $\overline{K}$, and the valuation ring of a place corresponds to those functions which are defined at that point. The discrete valuation of a place tells us to what degree a function vanishes or has a pole at the corresponding point. The divisor of a rational function then consists of all its zeroes minus all its poles, each counted with multiplicity.

The *Riemann-Roch space* of a divisor $D \in \mathrm{Div}(F)$ is defined by

$$\mathscr{L}(D) \coloneqq \{f \in F^* \mid (f) + D \geq 0\} \cup \{0\},$$

and we denote by $\ell(D)$ its dimension as a K-vector space. If $D = \sum_{P \in \mathscr{C}} n_P P$, the elements of $\mathscr{L}(D)$ are the functions which have zeroes of order at least $-n_P$ at all points P with $n_P < 0$, and are only allowed to have poles at points P with $n_P > 0$, of order at most n_P.

The *genus* of the function field F is defined as $g \coloneqq \max\{\deg(D) - \ell(D) + 1 \mid D \in \mathrm{Div}(F)\}$. It is the same as the genus of the corresponding algebraic curve. The following famous theorem about the dimension of the Riemann-Roch space is essential to this chapter:

Theorem 2.1 (Riemann-Roch)**.** *Let g be the genus of F, and let $D \in \mathrm{Div}(F)$ be a divisor. Then,*

$$\deg(D) + 1 - g \leq \ell(D) \leq \deg(D) + 1,$$

and the lower bound $\ell(D) = \deg(D) + 1 - g$ is achieved whenever $\deg(D) \geq 2g - 1$.

Recall that we defined the *holomorphy ring* of a nonempty set of places $\mathscr{S} \subsetneq \mathscr{C}$ as

$$H \coloneqq \bigcap_{P \in \mathscr{S}} \mathcal{O}_P.$$

In the point of view of curves, H is the ring of all rational functions that are defined on $\mathscr{S}$. We have a bijection between the set of places $\mathscr{S}$ and the maximal ideals of H given by $P \mapsto P \cap H \coloneqq P^H$.

When the base field $K = \mathbb{F}_q$ is finite, we can define the *zeta function* of the function field F as

$$\zeta_F(s) \coloneqq \prod_{P \in \mathscr{C}} \left(1 - \frac{1}{q^{s \deg(P)}}\right)^{-1},$$

for $s > 1$. Analogously, we define the zeta function of the holomorphy ring H corresponding to the set of places $\mathscr{S}$ as

$$\zeta_H(s) \coloneqq \prod_{P \in \mathscr{S}} \left(1 - \frac{1}{q^{s \deg(P)}}\right)^{-1}.$$

2.3 Density in Holomorphy Rings

We require a definition of density in holomorphy rings that is analogous to the definition for integers in Definition 1.3. Intuitively, the density of a set $L \subseteq H^m$ should give the probability that a randomly selected m-tuple over H lies in L. For this, we first need to cover H with a suitable increasing sequence — or net — of finite subsets, for each of which the proportion of elements lying in L is well-defined. The overall density of L is then the limit of these proportions over the net of subsets. Apparently, the most natural such cover is a net of Riemann-Roch spaces, which we will describe now.

First, let us give the definition of Moore-Smith convergence of nets (see [Kel55, Chapter 2]). A *directed set* is a set $\mathscr{I}$, endowed with a relation $\preceq$ that is reflexive, transitive, and such that for each pair of elements $a, b \in \mathscr{I}$, there is a common upper bound $c \succeq a, b$. A *net* is then a map $f : \mathscr{I} \longrightarrow X$ from a directed set to a topological space. We define its limit as follows: $\lim_{a \in \mathscr{I}} f(a) = x$ if and only if for each neighbourhood U of x, there exists a $b \in \mathscr{I}$ such that $f(a) \in U$ for all $a \succeq b$. We also define the limit superior as $\limsup_{a \in \mathscr{I}} f(a) \coloneqq \lim_{a \in \mathscr{I}} \sup_{b \succeq a} f(b)$, and likewise the limit inferior. Note that in the case $\mathscr{I} = \mathbb{N}$, we get the usual definition of limits of sequences.

Recall that H is the holomorphy ring of the set $\mathscr{S} \subsetneq \mathscr{C}$. Define

$$\mathscr{D} \coloneqq \{D \in \mathrm{Div}(F) \mid D \geq 0 \wedge \mathrm{supp}(D) \subseteq \mathscr{C} \setminus \mathscr{S}\},$$

the set of positive divisors supported away from $\mathscr{S}$. This is a directed set with the usual partial order on divisors. Note now that the elements of the Riemann-Roch space $\mathscr{L}(D)$ are allowed to have poles only outside of $\mathscr{S}$, which means that $\mathscr{L}(D) \subseteq H$. Conversely, for any $f \in H$, its pole divisor $D = (f)_\infty$ lies in $\mathscr{D}$, and since $f \in \mathscr{L}(D)$ we get

$$H = \bigcup_{D \in \mathscr{D}} \mathscr{L}(D).$$

This is a suitable covering for our definition. We can now define the *superior density* of a subset $L \subseteq H^m$ as

$$\overline{\mathbb{D}}(L) \coloneqq \limsup_{D \in \mathscr{D}} \frac{\left|L \cap \mathscr{L}(D)^m\right|}{\left|\mathscr{L}(D)^m\right|}, \tag{2.1}$$

and the *inferior density* as

$$\underline{\mathbb{D}}(L) \coloneqq \liminf_{D \in \mathscr{D}} \frac{\left|L \cap \mathscr{L}(D)^m\right|}{\left|\mathscr{L}(D)^m\right|}.$$

Moreover, whenever $\overline{\mathbb{D}}(L) = \underline{\mathbb{D}}(L)$, we call this value the *density* of L and denote it by

$$\mathbb{D}(L) \coloneqq \lim_{D \in \mathscr{D}} \frac{\left|L \cap \mathscr{L}(D)^m\right|}{\left|\mathscr{L}(D)^m\right|}.$$

In the case where $\mathscr{C} \setminus \mathscr{S}$ is finite, an analogous definition of density can also be found in [Poo03, Section 8].

2.4 The Density of Coprime Tuples

For a fixed positive integer m, we wish to study the set of coprime m-tuples of elements of the ring H. Let us denote this set by U:

$$U \coloneqq \{f = (f_1, \dots, f_m) \in H^m \mid I_f = H\},$$

where I_f denotes the ideal of H generated by the set $\{f_1, \dots, f_m\}$. Our first main result gives the density of this set:

Theorem 2.2. *The density of the set of coprime tuples of length $m \geq 2$ of the holomorphy ring H is $\mathscr{D}(U) = \frac{1}{\zeta_H(m)}$.*

Proof. The proof of this can be found in [MS16a]. We will not repeat it here, since it is a special case of Theorem 2.8 and the ideas for the proofs are similar. □

The reader should observe that in Theorem 2.2, both $\mathscr{S}$ and $\mathscr{C} \setminus \mathscr{S}$ could possibly be infinite and the result will still hold. Nevertheless, the density depends on the zeta function of the holomorphy ring, which may be hard to compute. First of all, notice that this is not the case when $\mathscr{S}$ is finite since under this condition ζ_H is a finite product. The following immediate corollary covers the case in which $\mathscr{C} \setminus \mathscr{S}$ is finite.

Corollary 2.3. *Let F be a function field, $\mathscr{S}$ a set of places of F and H the holomorphy ring of $\mathscr{S}$. Let $L_F(T)$ be the L-polynomial of F. Then*

$$\zeta_H(m) = \frac{L_F(q^{-m})}{(1 - q^{-m})(1 - q^{-m+1})} \prod_{P \in \mathscr{C} \setminus \mathscr{S}} \left(1 - \frac{1}{q^{\deg(P)m}}\right).$$

Proof. The corollary follows from Theorem 2.2, the definition of ζ_H and the expression of the zeta function of F in terms of the L-polynomial. □

Remark 2.4. Observe now that in the case where $\mathscr{C} \setminus \mathscr{S}$ is finite, the density of coprime m-tuples of H depends only on the following finite data: The degrees of the places in $\mathscr{C} \setminus \mathscr{S}$ and the L-polynomial of the function field, which again only depends on the $\mathbb{F}_{q^i}$-rational points of the curve associated to the function field for $i \in \{1, \dots, g\}$ (see for example [Sti09, Corollary 5.1.17]).

2.4.1 An example

Let $\mathrm{char}(\mathbb{F}_q) \neq 2, 3$ for simplicity. Let $a, b \in \mathbb{F}_q$ and $p(x, y) = y^2 - x^3 - ax - b$ be a polynomial defining an elliptic curve E over $\mathbb{F}_q$. Let us define

$$A(E) \coloneqq \mathbb{F}_q[x, y]/(p(x, y)).$$

Let $E(\mathbb{F}_q)$ denote the set of (projective) $\mathbb{F}_q$-rational points of E (i.e. the places of degree one of the function field of E).

Corollary 2.5. *The density of m-tuples of coprime elements of $A(E)$ is*

$$\mathbb{D}(U) = \frac{1 - q^{-m+1}}{1 + a_q q^{-m} + q^{-2m+1}}, \tag{2.2}$$

where $a_q = q + 1 - |E(\mathbb{F}_q)|$.

Proof. Observe that the zeta function of an elliptic curve is

$$\zeta_E(s) = \frac{1 + a_q q^{-s} + q^{-2s+1}}{(1 - q^{-s})(1 - q^{-s+1})}.$$

The holomorpy ring of E is $A(E) = \bigcap_{P \neq P_\infty} \mathcal{O}_P$, where P_∞ is the place at infinity of E. The result then follows from Theorem 2.2. □

Remark 2.6. The reader should notice that (2.2) depends only on the number of $\mathbb{F}_q$-rational points of E, since the genus of E equals one (see Remark 2.4). The probabilistic interpretation of Corollary 2.5 is the following: Select uniformly at random m elements of $A(E)$ of degree at most N, then the probability that they generate the unit ideal in $A(E)$ approaches $\frac{1-q^{-m+1}}{1+a_q q^{-m}+q^{-2m+1}}$ as $N \longrightarrow \infty$.

2.4.2 The case $F = \mathbb{F}_q(x)$

We will now show how the results [ST07, Theorem 1] and [GY13, Remark 4] about coprime m-tuples over $\mathbb{F}_q[x]$ fit in our framework.

Denote by P_∞ the place at infinity of the function field $\mathbb{F}_q(x)$. It is easy to see that the definition of density for $\mathbb{F}_q[x]$ given in [GY13; ST07] agrees with ours for $H = \mathbb{F}_q[x] = \bigcap_{P \neq P_\infty} \mathcal{O}_P$. Hence, we get [GY13, Remark 4] as a corollary to Theorem 2.2, while [ST07, Theorem 1] is simply the special case $m = 2$:

Corollary 2.7. *Let $m > 1$ be an integer. The density of coprime m-tuples over $\mathbb{F}_q[x]$ is*

$$\mathbb{D}(U) = 1 - \frac{1}{q^{m-1}}.$$

Proof. It is enough to notice that the zeta function of the function field $\mathbb{F}_q(x)$ (i.e. the zeta function of the projective line) is

$$\zeta_{\mathbb{F}_q(x)}(s) = \frac{1}{(1 - q^{-s})(1 - q^{-s+1})},$$

and then the zeta function of the holomorphy ring $\mathbb{F}_q[x] = \bigcap_{P \neq P_\infty} \mathcal{O}_P$ is

$$\zeta_{\mathbb{F}_q[x]}(s) = \frac{1}{1 - q^{-s+1}}.$$

The claim follows by inverting the expression above and evaluating at m. □

2.5 The Density of Unimodular Matrices

We now wish to study the $k \times m$-matrices over the holomorphy ring H which can be extended to $m \times m$ unimodular matrices, for some positive integers $k < m$. In other words, our goal is to compute the density of the set

$$U \coloneqq \left\{ A \in H^{k\times m} \;\middle|\; \exists A' \in H^{(m-k)\times m} : \left(\begin{smallmatrix} A \\ A' \end{smallmatrix}\right) \in \mathrm{GL}_m(H) \right\}.$$

Our main result gives the density of U in the case $2k - 1 \leq m$.

Theorem 2.8. *Let $2k - 1 \leq m$, then*

$$\mathbb{D}(U) = \prod_{j=m-k+1}^{m} \frac{1}{\zeta_H(j)}.$$

Here, the density of $k \times m$-matrices is defined like in Section 2.3 by considering them as km-tuples. Note that for $k = 1$, the $1 \times m$-matrix A lies in U if and only if its entries are coprime. Hence, this theorem is a generalization of Theorem 2.2.

The basic ingredients for the proof are the Riemann-Roch theorem and the Hasse-Weil bound. In any case, the main difficulty we will encounter is giving an estimate for the number of matrices having entries in $\mathscr{L}(D)$ whose reduction modulo a place P is not of full rank. This is provided in Lemma 2.12.

2.5.1 An Unconditional Upper Bound

On the way to our proof of Theorem 2.8, we first give an upper bound to the superior density $\overline{\mathbb{D}}(U)$. This bound does not require the condition $2k - 1 \leq m$, but holds whenever $k \leq m$.

Note first that the set U of $k \times m$-matrices which can be extended to unimodular matrices is exactly

$$U = \{A \in H^{k\times m} \mid I_A = H\},$$

where I_A is the ideal generated by the $k \times k$ minors of A (see for example [GMR81]). The condition $I_A = H$ holds if and only if $I_A + Q^H = H$ for every place $Q \in \mathcal{S}$, where $Q^H = Q \cap H$, since these Q^H range over all maximal ideals of H.

In order to study the density of U, we proceed by first restricting our attention to only finitely many places. Hence, for a finite subset $\mathcal{Q} \subseteq \mathcal{S}$, we define

$$U_{\mathcal{Q}} \coloneqq \{A \in H^{k\times m} \mid I_A + Q^H = H \text{ for all } Q \in \mathcal{Q}\}.$$

As discussed above, U is the intersection of the $U_{\mathcal{Q}}$, with $\mathcal{Q}$ ranging over all finite subsets of $\mathcal{S}$. The following lemma gives the density of $U_{\mathcal{Q}}$. It is very similar to a part of the main proof in [MS16a].

Lemma 2.9. *Let $\mathcal{Q} = \{Q_1, \dots, Q_n\} \subseteq \mathcal{S}$ be a finite subset of places. The set $U_{\mathcal{Q}} \subseteq H^{k\times m}$ has density*

$$\mathbb{D}(U_{\mathcal{Q}}) = \prod_{i=1}^{n}\prod_{j=0}^{k-1}(1 - q^{(j-m)\deg Q_i}).$$

Proof. A matrix $A \in H^{k\times m}$ lies in $U_{\mathcal{Q}}$ if and only if for each $i \in \{1, \dots, n\}$, the ideal I_A is not contained in Q_i^H, which is to say that the image of A in $(H/Q_i^H)^{k\times m} \cong \mathbb{F}_{q^{\deg Q_i}}^{k\times m}$ has some nonzero $k \times k$ minors, i.e. has full rank.

Consider now the projection

$$\phi : H \longrightarrow H/(Q_1^H \cdots Q_n^H) \cong \prod_{i=1}^{n} H/Q_i^H \cong \prod_{i=1}^{n} \mathbb{F}_{q^{\deg Q_i}},$$

where the first isomorphism is by the Chinese remainder theorem. The number of matrices in $\prod_{i=1}^{n} \mathbb{F}_{q^{\deg Q_i}}^{k\times m}$ with full rank in each component is

$$\prod_{i=1}^{n}\prod_{j=0}^{k-1}(q^{m\deg Q_i} - q^{j\deg Q_i}) \tag{2.3}$$

by a simple counting argument. These form exactly the image of $U_{\mathcal{Q}}$ under $\phi^{k\times m}$.

Consider now a divisor $D \in \mathcal{D}$. We wish to count the number of matrices in $U_{\mathcal{Q}} \cap \mathcal{L}(D)^{k\times m}$. First, we will show that ϕ maps $\mathcal{L}(D)$ surjectively onto $\prod_{i=1}^{n} \mathbb{F}_{q^{\deg Q_i}}$ if $\deg D$ is large enough.

For this, note that the image of $\mathcal{L}(D)$ under ϕ is $\mathcal{L}(D)/(\mathcal{L}(D) \cap (Q_1^H \cdots Q_n^H))$. The space

$$\mathcal{L}(D) \cap (Q_1^H \cdots Q_n^H) = \mathcal{L}(D) \cap Q_1 \cap \cdots \cap Q_n$$

consists of all elements in $\mathcal{L}(D)$ with a root at each Q_i, so it is equal to $\mathcal{L}(D - \sum_{i=1}^{n} Q_i)$. Hence, its dimension as an $\mathbb{F}_q$-vector space is $\ell(D - \sum_{i=1}^{n} Q_i)$, which by the Riemann-Roch theorem is equal to $\deg D - \sum_{i=1}^{n} \deg Q_i + 1 - g$ if $\deg D$ is large enough. On the other hand, the dimension of $\mathcal{L}(D)$ is then $\ell(D) = \deg D + 1 - g$, and so the image of $\mathcal{L}(D)$ under ϕ has dimension $\sum_{i=1}^{n} \deg Q_i$, the same as $\prod_{i=1}^{n} \mathbb{F}_{q^{\deg Q_i}}$.

This argument also gives us the dimension of the kernel of ϕ restricted to $\mathcal{L}(D)$, which is $\ell(D - \sum_{i=1}^{n} Q_i)$.

From this together with (2.3), we can count the number of matrices in $U_{\mathcal{Q}} \cap \mathcal{L}(D)^{k\times m}$, and we get that

$$\begin{aligned}\frac{\left|U_{\mathcal{Q}} \cap \mathcal{L}(D)^{k\times m}\right|}{\left|\mathcal{L}(D)^{k\times m}\right|} &= q^{km(\ell(D-\sum_{i=1}^{n} Q_i)-\ell(D))} \cdot \prod_{i=1}^{n}\prod_{j=0}^{k-1}(q^{m\deg Q_i} - q^{j\deg Q_i}) \\ &= \prod_{i=1}^{n}\prod_{j=0}^{k-1}(1 - q^{(j-m)\deg Q_i})\end{aligned}$$

if $\deg D$ is large enough. Taking the limit over all $D \in \mathcal{D}$, we get the claim. □

Proposition 2.10. *We have an upper bound for the superior density of U given by*

$$\overline{\mathbb{D}}(U) \leq \prod_{j=0}^{k-1} \prod_{Q \in \mathcal{S}} (1 - q^{(j-m)\deg Q}).$$

Proof. For each finite subset $\mathcal{Q} \subseteq \mathcal{S}$, we have that $U \subseteq U_{\mathcal{Q}}$, and hence by Lemma 2.9

$$\overline{\mathbb{D}}(U) \leq \mathbb{D}(U_{\mathcal{Q}}) = \prod_{j=0}^{k-1} \prod_{Q \in \mathcal{Q}} (1 - q^{(j-m)\deg Q}).$$

By taking the Moore-Smith limit over all such finite sets $\mathcal{Q}$, we get our claim. □

Remark 2.11. The possibly infinite product $\prod_{Q \in \mathcal{S}} (1 - q^{(j-m)\deg Q})$ in Proposition 2.10 converges absolutely if $j < m - 1$, since it is a subproduct of the reciprocal of the zeta function of F. If $k = m$, we get the term $\prod_{Q \in \mathcal{S}} (1 - q^{-\deg Q})$, which corresponds to the pole of the zeta function. Nevertheless, since each term is strictly between 0 and 1, the product makes sense, though it may diverge to zero.

2.5.2 The Case $2k - 1 \leq m$

In the case where $2k - 1 \leq m$, we can prove that the upper bound of Proposition 2.10 is in fact exact. For this, we will need the following lemma.

Lemma 2.12. *Let n be an integer and S be a nonempty subset of $\mathbb{F}_{q^n}$. Let N be the set of $k \times m$ full rank matrices having entries in S. Then, independently of n, we have that*

$$|N| \geq \prod_{i=0}^{k-1} (|S|^m - |S|^i).$$

Under these assumptions, this bound is the best possible.

Proof. Let us bound $|N|$ by progressively counting the rows which can occur for a matrix in N. The first row $(v_{1,1}, \ldots, v_{1,m})$ of a matrix A in N can be fixed in at least $|S|^m - 1$ ways: Everything but the zero row can be extended to be invertible. The second row can be fixed in $|S|^m - |B_1|$ ways, where B_1 are the $\mathbb{F}_{q^n}$-multiples of the first row which have all entries in S. Since the first row is nonzero, there exists $v_{1,j}$ different from zero. By the obvious inclusion of sets, we can now bound $|B_1|$ from above with the number of $\mathbb{F}_{q^n}$-multiples of the first row which have the j-th component in S. These are in bijection with the $a \in \mathbb{F}_{q^n}$ for which $a v_{1,j} \in S$, of which there are exactly $|S|$. Hence, $|B_1| \leq |S|$.

We can iterate this procedure as follows: The i-th row $(v_{i,1}, \ldots, v_{i,m})$ can be chosen in $|S|^m - |B_{i-1}|$ ways, where B_{i-1} are the rows in S^m which are $\mathbb{F}_{q^n}$-linearly dependent on the first $i - 1$ rows. By construction, the first $i - 1$ rows are linearly independent, which implies that there exists a full rank $(i-1) \times (i-1)$ submatrix K. Let $j_1, \ldots, j_{i-1} \in \{1, \ldots, m\}$ be the indices corresponding to the columns of the full

rank submatrix. As in the simpler case before, B_{i-1} is contained in the set of rows which are $\mathbb{F}_{q^n}$-linearly dependent on the first $i-1$ and for which the components at indices $j_1, \dots, j_{i-1}$ lie in S. It is easily seen that the rows satisfying this weaker condition are in bijection with the vectors $w \in \mathbb{F}_{q^n}^{i-1}$ such that $wK \in S^{i-1}$, so there are exactly $|K^{-1}S^{i-1}| = |S|^{i-1}$ of them, and $|B_{i-1}| \leq |S|^{i-1}$.

This gives us the bound of the lemma. The reader should now observe that this bound is uniform in n and also that it is the best possible general bound, since it is attained when S is a subfield of $\mathbb{F}_{q^n}$. □

With this, we can compute the inferior density of U.

Proposition 2.13. *If $2k-1 \leq m$, we have for the inferior density of U*

$$\underline{\mathbb{D}}(U) = \prod_{j=0}^{k-1} \prod_{Q \in \mathcal{S}} (1 - q^{(j-m)\deg Q}).$$

Proof. Fix a finite set of places $\mathcal{Q} \subseteq \mathcal{S}$. Having $U \subseteq U_{\mathcal{Q}}$, it is easily seen that (see also [DM16, Lemma 4])

$$\underline{\mathbb{D}}(U) = \mathbb{D}(U_{\mathcal{Q}}) - \overline{\mathbb{D}}(U_{\mathcal{Q}} \setminus U).$$

If we prove that $\overline{\mathbb{D}}(U_{\mathcal{Q}} \setminus U)$ tends to zero, the claim follows from Lemma 2.9 by taking the limit over all finite subsets $\mathcal{Q} \subseteq \mathcal{S}$.

To show this, we first fix the set of places $\mathcal{Q}$, then a divisor $D \in \mathcal{D}$ depending on $\mathcal{Q}$. We can now write:

$$\begin{aligned}(U_{\mathcal{Q}} \setminus U) \cap \mathcal{L}(D)^{k\times m} &\subseteq \{A \in \mathcal{L}(D)^{k\times m} \mid I_A \subseteq P^H \text{ for some } P \in \mathcal{S} \setminus \mathcal{Q}\} \\ &= \bigcup_{P \in \mathcal{S}\setminus\mathcal{Q}} \{A \in \mathcal{L}(D)^{k\times m} \mid I_A \subseteq P^H\}.\end{aligned}$$

In this union, we can limit ourselves to P of degree at most $k \deg D$: Indeed, suppose we have $\deg P \geq k \deg D$ and $A \in \mathcal{L}(D)^{k\times m}$ with $I_A \subseteq P^H$. Since all entries of A are in $\mathcal{L}(D)$ and the $k \times k$ minors are degree k polynomials in these, we see that each such minor lies in $\mathcal{L}(kD)$. Because $I_A \subseteq P^H$, they furthermore lie in $\mathcal{L}(kD - P)$, which is trivial when $\deg(kD - P) = k \deg D - \deg P < 0$. The ideal I_A is generated by these minors and is hence (0). It follows that $I_A \subseteq P'^H$ for every place P', and so A is already contained in the following restricted union (assuming D is chosen large enough that there is at least one $P' \in \mathcal{S} \setminus \mathcal{Q}$ of degree less than $k \deg D$):

$$(U_{\mathcal{Q}} \setminus U) \cap \mathcal{L}(D)^{k\times m} \subseteq \bigcup_{\substack{P \in \mathcal{S}\setminus\mathcal{Q} \\ \deg P \leq k \deg D}} \{A \in \mathcal{L}(D)^{k\times m} \mid I_A \subseteq P^H\}.$$

We write $S_P \coloneqq \{A \in \mathcal{L}(D)^{k\times m} \mid I_A \subseteq P^H\}$.

With this, we can now give the following estimate:

$$\frac{\left|(U_{\mathbb{Q}} \setminus U) \cap \mathscr{L}(D)^{k\times m}\right|}{\left|\mathscr{L}(D)^{k\times m}\right|} \leq \sum_{\substack{P\in\mathcal{S}\setminus\mathbb{Q}\\ \deg P\leq k\deg D}} \frac{\left|S_P\right|}{\left|\mathscr{L}(D)^{k\times m}\right|}.$$

We concentrate on the individual summands. Consider as in Lemma 2.9 the projection

$$\phi:\ H \longrightarrow H/P^H \cong \mathbb{F}_{q^{\deg P}}$$

and its restriction to $\mathscr{L}(D)$, as an $\mathbb{F}_q$-linear map. Let d be the dimension of $\phi(\mathscr{L}(D))$. We easily see that $\phi^{k\times m}(S_P)$ is the set of matrices with entries in $\phi(\mathscr{L}(D))$ not of full rank, which by Lemma 2.12 has cardinality

$$|\phi^{k\times m}(S_P)| \leq q^{mkd} - \prod_{i=0}^{k-1}(q^{md} - q^{id}).$$

Furthermore, the kernel of $\phi|_{\mathscr{L}(D)}$ is $\mathscr{L}(D-P)$, so we get

$$\begin{aligned}|S_P| &\leq q^{mk\ell(D-P)}\cdot\left(q^{mkd} - \prod_{i=0}^{k-1}(q^{md}-q^{id})\right)\\ &\leq q^{mk\ell(D-P)}\cdot Cq^{(k-1)(m+1)d},\end{aligned}$$

where C is a constant depending only on k ($C = 2^k$ works). We have $d = \ell(D) - \ell(D-P)$ by the rank-nullity theorem, so we can write

$$\begin{aligned}\frac{\left|S_P\right|}{\left|\mathscr{L}(D)^{k\times m}\right|} &\leq q^{mk(\ell(D-P)-\ell(D))}\cdot Cq^{(k-1)(m+1)d}\\ &= q^{-mkd}\cdot Cq^{(k-1)(m+1)d}\\ &= Cq^{(k-m-1)d}\\ &\leq Cq^{-kd},\end{aligned} \tag{2.4}$$

the last following from the assumption $2k - 1 \leq m$. We now look at two cases separately.

Case 1: $\deg D < \deg P \leq k\deg D$. In this case, we see that $\ell(D-P) = 0$ and $d = \ell(D)$. The Hasse-Weil bound says that the number of places of F of degree at most r is asymptotically equal to $\frac{q+1}{q}\frac{q^r}{r}$. With this and (2.4), we can estimate for large D the sum

$$\sum_{\substack{P\in\mathcal{S}\setminus\mathbb{Q}\\ \deg D<\deg P\leq k\deg D}} \frac{\left|S_P\right|}{\left|\mathscr{L}(D)^{k\times m}\right|} \leq C'\cdot\frac{q^{k\deg D}}{k\deg D}\cdot q^{-k\ell(D)},$$

for some new constant $C' = C \cdot (\frac{q+1}{q} + 1)$. Since $\ell(D) = \deg D + 1 - g$ for large D, we can rewrite the above to

$$\frac{C'}{k \deg D} \cdot q^{-k(1-g)}$$

which goes to zero upon taking the limit in D.

Case 2: $\deg P \leq \deg D$. In this case, we note that $\ell(D) \geq \deg D + 1 - g$ and $\ell(D-P) \leq \deg(D-P)+1$ by Riemann-Roch, and so $d = \ell(D) - \ell(D-P) \geq \deg P - g$. Using this with (2.4), we can estimate the sum

$$\sum_{\substack{P \in \mathcal{S} \setminus \mathcal{Q} \\ \deg P \leq \deg D}} \frac{\left|S_P\right|}{\left|\mathscr{L}(D)^{k \times m}\right|} \leq \sum_{\substack{P \in \mathcal{S} \setminus \mathcal{Q} \\ \deg P \leq \deg D}} C' \cdot q^{-k \deg P},$$

for $C' = Cq^{kg}$. Taking the limit in D, we get

$$\sum_{P \in \mathcal{S} \setminus \mathcal{Q}} C' \cdot q^{-k \deg P}.$$

Up to the factor C', this is the tail of a subseries of the zeta function of F evaluated at k, which converges absolutely for $k > 1$. For $k = 1$, $m > 1$, note that the inequality $2k - 1 \leq m$ is not sharp, so the proof still works with a slight modification to the estimate (2.4).

Putting together all we have done so far, we see that

$$\begin{aligned}
\overline{\mathbb{D}}(U_{\mathcal{Q}} \setminus U) &= \limsup_{D \in \mathscr{D}} \frac{\left|(U_{\mathcal{Q}} \setminus U) \cap \mathscr{L}(D)^{k \times m}\right|}{\left|\mathscr{L}(D)^{k \times m}\right|} \\
&\leq \limsup_{D \in \mathscr{D}} \left(\sum_{\substack{P \in \mathcal{S} \setminus \mathcal{Q} \\ \deg D < \deg P \leq k \deg D}} \frac{\left|S_P\right|}{\left|\mathscr{L}(D)^{k \times m}\right|} + \sum_{\substack{P \in \mathcal{S} \setminus \mathcal{Q} \\ \deg P \leq \deg D}} \frac{\left|S_P\right|}{\left|\mathscr{L}(D)^{k \times m}\right|} \right) \\
&\leq 0 + \sum_{P \in \mathcal{S} \setminus \mathcal{Q}} C' \cdot q^{-k \deg P}.
\end{aligned}$$

This final sum can be seen as the tail of an absolutely convergent series, and hence it converges to zero as $\mathcal{Q}$ grows. We conclude that

$$\lim_{\text{finite } \mathcal{Q} \subseteq \mathcal{S}} \overline{\mathbb{D}}(U_{\mathcal{Q}} \setminus U) = 0,$$

from which the proposition follows. □

The main result, Theorem 2.8, now follows immediately from Propositions 2.10 and 2.13.

Remark 2.14. We tried to keep the tools used as elementary as possible. Nevertheless, the reader who is willing to pursue a more sophisticated approach (based on p-adic analysis) can refer to a function field version of [PS99, Lemma 20]. Still, a part of Lemma 2.9 will be needed and Lemma 2.12 will still be entirely necessary to satisfy the condition of [PS99, Lemma 20].

Remark 2.15. It is worth observing that in [GY13], this result has been presented for the special case of polynomial rings, which can be regarded as holomorphy rings of the rational function field, as in Section 2.4.2. Unfortunately, the proof of [GY13] is not correct as it is presented (see the exchange of lim sup and an infinite sum in the proof of [GY13, Theorem 1]) and needs a fix, which we were able to perform in the case $2k - 1 \leq m$. It would be of great interest if one could adjust this proof to work in the most general case (i.e. $k < m$), which we conjecture to be true at least in the case of the polynomial ring.

Chapter 3

Density of Shifted Eisenstein Polynomials

3.1 Introduction

The Eisenstein irreducibility criterion [Eis50; Sch46] is a very convenient tool to establish that a polynomial in $\mathbb{Z}[x]$ is irreducible.[1]

Definition 3.1. Let R be an integral domain and $R[x]$ be the ring of polynomials with coefficients in R. We say that $f(x) = \sum_{i=0}^{n} \alpha_i x^i \in R[x]$ of degree n is *Eisenstein with respect to a prime ideal* P or *P-Eisenstein* if

- $\alpha_n \notin P$,
- $\alpha_i \in P$ for all $i \in \{0, \dots, n-1\}$,
- $\alpha_0 \notin P^2$.

We say that $f(x)$ is *Eisenstein* if it is Eisenstein with respect to some prime ideal P.

The Eisenstein criterion now says that any Eisenstein polynomial is irreducible in $F[x]$, where F is the field of fractions of R. In this chapter, we will only consider the case of the ring of integers $R = \mathbb{Z}$ and the rings of p-adic integers $R = \mathbb{Z}_p$.

It is a well understood fact that the density of irreducible polynomials of fixed degree n among all the polynomials of degree n is equal to one. The question which naturally arises is the following:

Question 3.2. What is the natural density of degree n polynomials which are Eisenstein?

[1] In this chapter, when we say that a polynomial in $\mathbb{Z}[x]$ is irreducible, we mean that it is irreducible as an element of $\mathbb{Q}[x]$.

More informally, how likely is it that checking whether a random polynomial is irreducible using only the Eisenstein irreducibility criterion leads to success? In [Dub03; HS13], the authors deal with Eisenstein polynomials of fixed degree with coefficients over $\mathbb{Z}$. They provide a complete answer to the above question in the case of monic (see [Dub03, Theorem 1] and [HS13, Theorem 1]) and non-monic (see [HS13, Theorem 2]) Eisenstein polynomials.

We will specialize to the case of non-monic Eisenstein polynomials, since the proofs and methods can be easily adapted from one case to the other. In [HS13], the authors consider the set of polynomials of degree at most n having integer coefficients bounded in absolute value by B (the *height* of a polynomial) and give a sharp estimate for the number $\rho(B)$ of polynomials which are irreducible by the Eisenstein criterion. The natural density of degree n Eisenstein polynomials is then the limit of the sequence $\rho(B)/(2B)^{n+1}$, which turns out to be

$$1 - \prod_{p \text{ prime}} \left(1 - \frac{(p-1)^2}{p^{n+2}}\right).$$

This fully answers Question 3.2.

Clearly, a polynomial $f(x)$ is irreducible if and only if $f(x+i)$ is irreducible for any and hence all $i \in \mathbb{Z}$. Using this simple observation, one could check irreducibility by trying to use the Eisenstein criterion for many i. How likely is it that this procedure works? More formally,

Question 3.3. What is the natural density of degree n polynomials $f(x)$ for which $f(x+i)$ is irreducible by the Eisenstein criterion for some integer shift i?

In [HS14], Heyman and Shparlinski address this question, and give a lower bound on this density. Nevertheless, the question regarding the exact density remained open. We managed to give an exact solution to this question in our paper [MS16b], which is the basis of this chapter.

In Section 3.2, we give the complete solution to Question 3.3 using a local to global principle for densities [PS99, Lemma 20]. Using similar methods, we also provide a solution to the question appearing in [HS14, Section 7] about *affine* Eisenstein polynomials in Section 3.3.

Our proofs are also supported by Monte Carlo experiments which we provide in Section 3.4.

3.2 Shifted Eisenstein Polynomials

In this section, we determine the density of degree n polynomials $f(x) \in \mathbb{Z}[x]$ such that $f(x+i)$ is Eisenstein for some shift $i \in \mathbb{Z}$.

First, we need to define what density even means for subsets of $\mathbb{Z}[x]_{\leq n}$, i.e. of the set of polynomials of degree at most n. We do this by simply identifying the module $R[x]_{\leq n}$ with R^{n+1} via the standard basis $\{1, x, \ldots, x^n\}$, for any ring R. Then, we can apply Definition 1.3.

Let σ be the shift map defined by

$$\begin{aligned} \sigma : \mathbb{Z}^{n+1} &\longrightarrow \mathbb{Z}^{n+1} \\ f(x) &\longmapsto f(x+1). \end{aligned}$$

It is easy to see that σ is a linear map with determinant one. Similarly, we get a determinant one map over $\mathbb{Z}_p$ for any p, which we will also denote by σ.

Definition 3.4. Let $E \subseteq \mathbb{Z}^{n+1}$ be the set of degree n Eisenstein polynomials over the integers. Let E_p be the set of degree n Eisenstein polynomials over $\mathbb{Z}_p$.

Definition 3.5. Let $\overline{E} \subseteq \mathbb{Z}^{n+1}$ be the set of degree n polynomials which are Eisenstein after applying some shift $i \in \mathbb{Z}$:

$$\overline{E} \coloneqq \{ f(x) \in \mathbb{Z}^{n+1} \mid f(x+i) \in E \text{ for some } i \in \mathbb{Z} \}.$$

We call these polynomials *shifted Eisenstein.*

Remark 3.6. When we consider the set $\overline{E} \subseteq \mathbb{Z}^{n+1}$, we are computing the density of shifted (and later affine) Eisenstein polynomials of degree *exactly n* among polynomials of degree *at most n*. Nevertheless it is easy to see that the density of shifted (and also affine, see Remark 3.16) Eisenstein polynomials of degree less or equal than n is the same.

In order to compute the density of $\overline{E}$, we need to consider each prime p separately. We do this by working over the p-adic integers.

Definition 3.7. Let $\overline{E}_p \subseteq \mathbb{Z}_p^{n+1}$ be the set of degree n polynomials of $\mathbb{Z}_p[x]$ which are Eisenstein after applying some shift $i \in \mathbb{Z}_p$:

$$\overline{E}_p \coloneqq \{ f(x) \in \mathbb{Z}_p^{n+1} \mid f(x+i) \in E_p \text{ for some } i \in \mathbb{Z}_p \}.$$

We also call these polynomials shifted Eisenstein, since it will always be clear from the context to which set we are referring.

Notice that $\overline{E}_p \cap \mathbb{Z}^{n+1}$ are exactly the polynomials of $\mathbb{Z}[x]$ of degree n which are shifted p-Eisenstein.

Lemma 3.8. *If $n \geq 2$ and $f(x) \in \mathbb{Z}_p^{n+1}$ is shifted Eisenstein, then it is so with respect to exactly one rational integer shift $i \in \{0, \dots, p-1\}$. In other words,*

$$\overline{E}_p = \bigsqcup_{i=0}^{p-1} \sigma^{-i} E_p.$$

Proof. We clearly have

$$\bigcup_{i=0}^{p-1} \sigma^{-i} E_p \subseteq \overline{E}_p.$$

The other inclusion is easy but not completely trivial.

Let $f(x) = \sum_{i=0}^{n} \alpha_i x^i \in E_p$ and $k \in \mathbb{Z}_p$. We first show that $f(x + kp)$ is also Eisenstein: Clearly $f(x) = f(x + kp)$ in $\mathbb{Z}_p/p\mathbb{Z}_p$, so the only condition which one has to check is that the coefficient of the term of degree zero of $f(x + kp)$ is not in $p^2\mathbb{Z}_p$. This coefficient is in fact $f(kp) = \alpha_0 + \alpha_1 kp + \sum_{i=2}^{n} \alpha_i k^i p^i$. Modulo $p^2\mathbb{Z}_p$ we have that

- $\alpha_i k^i p^i$ is congruent to zero for $i \geq 2$,
- $\alpha_1 kp$ is congruent to zero since α_1 is in $p\mathbb{Z}_p$,
- α_0 is not congruent to zero since the polynomial $f(x)$ is Eisenstein,

from which it follows that the polynomial $f(x + kp)$ is Eisenstein in $\mathbb{Z}_p[x]$.

Let now $f(x) \in \overline{E}_p$, then $f(x + u)$ is Eisenstein for some $u \in \mathbb{Z}_p$. The inclusion

$$\overline{E}_p \subseteq \bigcup_{i=0}^{p-1} \sigma^{-i} E_p$$

will follow if we show that we can select u in $\{0, \dots, p-1\}$. Write $u = kp + i$ with $i \in \{0, \dots, p-1\}$ and $k \in \mathbb{Z}_p$. Using what we proved above, we see that $f(x + u - kp) = f(x + i)$ is Eisenstein, and the inclusion follows.

We now show that the union is disjoint, i.e. $\sigma^{-i} E_p \cap \sigma^{-j} E_p = \emptyset$ for any $i, j \in \{0, \dots, p-1\}$ and $i \neq j$. Without loss of generality, we can assume $i > j$. Then

$$\sigma^{-i} E_p \cap \sigma^{-j} E_p = \emptyset \iff E_p \cap \sigma^{i-j} E_p = \emptyset.$$

Let $t \coloneqq j - i$ and $\sum_{i=0}^{n} \alpha_k x^k = f(x) \in E_p$, then the coefficient of the degree zero term of $f(x + t)$ is $f(t) = \alpha_n t^n + \sum_{k=0}^{n-1} \alpha_k t^k$. Now, the reduction of α_k modulo p is zero for any $k \leq n - 1$ and α_n and t are invertible modulo p, so $f(x + t)$ is not Eisenstein. □

Let μ_p be the Haar measure on $\mathbb{Z}_p^{n+1}$ normalized to have total mass 1, and μ_∞ the Lebesgue measure on $\mathbb{R}^{n+1}$. (For basics on the Haar measure on the p-adic numbers, we refer to [Rob00].)

Lemma 3.9. *In the above notation we have for $n \geq 2$*

$$\mu_p(\overline{E}_p) = \frac{(p-1)^2}{p^{n+1}}.$$

Proof. Since σ^{-1} has determinant one, it does not change the p-adic volumes. Therefore, by Lemma 3.8, one has $\mu_p(\overline{E}_p) = p \cdot \mu_p(E_p)$. It is easy to compute the measure $\mu_p(E_p)$ by writing $E_p = (p\mathbb{Z}_p \setminus p^2\mathbb{Z}_p) \times (p\mathbb{Z}_p)^{n-1} \times (\mathbb{Z}_p \setminus p\mathbb{Z}_p)$. □

In order to obtain the density $\mathbb{D}(\overline{E})$ from the local data $\{\mu_p(\overline{E}_p)\}_p$, we will use the following lemma [PS99, Lemma 20].

Lemma 3.10. *Suppose $U_\infty \subseteq \mathbb{R}^d$ is such that $\mathbb{R}^+ \cdot U_\infty = U_\infty$, $\mu_\infty(\partial U_\infty) = 0$. Let $U_\infty^1 = U_\infty \cap [-1,1]^d$ and $s_\infty = \mu_\infty(U_\infty^1)$. Let $U_p \subseteq \mathbb{Z}_p^d$, $\mu_p(\partial U_p) = 0$ and $s_p = \mu_p(U_p)$ for each prime p. Let $M_\mathbb{Q}$ be the set of places of $\mathbb{Q}$. Moreover, suppose that*

$$\lim_{M\to\infty} \overline{\mathbb{D}}(\{a \in \mathbb{Z}^d \mid a \in U_p \text{ for some finite prime } p \text{ greater than } M\}) = 0. \tag{3.1}$$

Let $P : \mathbb{Z}^d \longrightarrow 2^{M_\mathbb{Q}}$ be defined as $P(a) = \{v \in M_\mathbb{Q} \mid a \in U_v\}$. Then we have:

(a) *$\sum_v s_v$ converges.*

(b) *For any $T \subseteq 2^{M_\mathbb{Q}}$, $\nu(T) \coloneqq \mathbb{D}(P^{-1}(T))$ exists and defines a measure on $2^{M_\mathbb{Q}}$, which is concentrated at the finite subsets of $M_\mathbb{Q}$.*

(c) *Let S be a finite subset of $M_\mathbb{Q}$, then*

$$\nu(\{S\}) = \prod_{v\in S} s_v \prod_{v\notin S}(1 - s_v).$$

Proof. For the proof, see [PS99, Lemma 20]. □

After showing that condition (3.1) applies, we can use Lemma 3.10 to determine the density of shifted Eisenstein polynomials over the integers.

Theorem 3.11. *Let $n \geq 3$. The density of shifted Eisenstein polynomials of degree n is*

$$\mathbb{D}(\overline{E}) = 1 - \prod_{p \text{ prime}} \left(1 - \frac{(p-1)^2}{p^{n+1}}\right). \tag{3.2}$$

Proof. Set $U_p = \overline{E}_p$ for all p and $U_\infty = \emptyset$. The conditions $\mu_p(\partial U_p) = 0$ hold since U_p is both closed and open. Notice that in the notation of Lemma 3.10 we have that $P^{-1}(\{\emptyset\})$ equals the complement of $\overline{E}$. Therefore, if condition (3.1) is verified, we get the claim:

$$\mathbb{D}(\overline{E}) = 1 - \prod_{p \text{ prime}} \big(1 - s_p\big) = 1 - \prod_{p \text{ prime}} \left(1 - \frac{(p-1)^2}{p^{n+1}}\right).$$

Let us now show that the condition indeed holds:

$$\begin{aligned} &\lim_{M\to\infty} \overline{\mathbb{D}}(\{a \in \mathbb{Z}^{n+1} \mid a \in \overline{E}_p \text{ for some finite prime } p \text{ greater than } M\}) \\ &\qquad = \lim_{M\to\infty} \limsup_{B\to\infty} \frac{\left|\bigcup_{p>M} \overline{E}_p \cap [-B, B[^{n+1}\right|}{(2B)^{n+1}}. \end{aligned} \tag{3.3}$$

We have $\overline{E}_p \cap [-B, B[^{n+1} = \emptyset$ for $p > CB^2$, where C is a constant depending only on the degree n. One can see that using the following argument: Let $f(x)$ be a polynomial in $[-B, B[^{n+1}$ for which $f(x+i)$ is Eisenstein, then [HS14, Lemma 3.1]

$$p^{n-1} \nmid \operatorname{disc}(f(x+i)) = \operatorname{disc}(f(x)) \neq 0.$$

Now, the discriminant of $f(x)$ is a polynomial of degree $2n-2$ in the coefficients, whence

$$p^{n-1} \leq \operatorname{disc}(f(x)) \leq DB^{2n-2}$$

for some constant D depending only on n. Therefore, for $C = D^{1/(n-1)}$, we have $p \leq CB^2$. Thus, we have just shown that for fixed B, the union in (3.3) is finite, and we can bound it by

$$\begin{aligned} &\lim_{M\to\infty} \limsup_{B\to\infty} \frac{\left|\bigcup_{CB^2>p>M} \overline{E}_p \cap [-B,B[^{n+1}\right|}{(2B)^{n+1}} \\ &\leq \lim_{M\to\infty} \limsup_{B\to\infty} \sum_{CB^2>p>M} \frac{\left|\overline{E}_p \cap [-B,B[^{n+1}\right|}{(2B)^{n+1}}. \end{aligned} \tag{3.4}$$

Given the order of the limits, we can fix the following setting: $M > n$ and $B > M$. Now let us bound $\left|\overline{E}_p \cap [-B,B[^{n+1}\right|$ in the following two cases:

(a) $2B < p$: In this case, we can consider $[-B,B[^{n+1}$ as a subset of $\mathbb{F}_p^{n+1}$ without losing any information. The reader should notice that modulo p, the elements of $\overline{E}_p$ have a multiple root of order n at some $i \in \mathbb{F}_p$. Now, the key observation is the following: The reduction modulo p of the polynomials in $[-B,B[^{n+1} \cap \overline{E}_p$ is contained in the set

$$S_p \coloneqq \{a(x-i)^n \mid a \in [-B,B[\setminus \{0\} \text{ and } -nai \in [-B,B[\}.$$

This represents the condition that the degree n and $n-1$ coefficients live in $[-B,B[$:

$$[-B,B[^{n+1} \cap \overline{E}_p \subseteq S_p.$$

Observe now that $|S_p| = (2B-1)2B \leq (2B)^2$, since n and a are invertible modulo p (recall $p > M > n$). We conclude that

$$\left|[-B,B[^{n+1} \cap \overline{E}_p\right| \leq |S_p| \leq (2B)^2.$$

Notice that this bound is uniform in p.

(b) $2B \geq p$: In this case, the bound is more natural. Consider the projection map

$$\pi : \mathbb{Z}^{n+1} \longrightarrow \mathbb{F}_p^{n+1}$$

and the shift map modulo p

$$\begin{aligned} \sigma^{-1} : \mathbb{F}_p^{n+1} &\longrightarrow \mathbb{F}_p^{n+1} \\ f(x) &\longmapsto f(x-1). \end{aligned}$$

Consider the sets of polynomials $L_p \coloneqq \{ax^n \mid a \in \mathbb{F}_p^*\}$ and

$$\overline{L}_p = \bigcup_{i=0}^{p-1} \sigma^{-i} L_p. \tag{3.5}$$

We have $|\overline{L}_p| \leq p^2$.

Notice that

$$\pi([-B, B[^{n+1} \cap \overline{E}_p) \subseteq \overline{L}_p. \tag{3.6}$$

At this step, we observe that the projection is at most $\lceil 2B/p \rceil^{n+1}$ to one, therefore we can bound $\left|[-B, B[^{n+1} \cap \overline{E}_p\right|$ using the projection map and condition (3.6):

$$\left|[-B, B[^{n+1} \cap \overline{E}_p\right| \leq |\overline{L}_p| \cdot \lceil 2B/p \rceil^{n+1} \leq p^2 \left(\frac{2B}{p} + 1\right)^{n+1} \leq p^2 \left(\frac{4B}{p}\right)^{n+1},$$

where the last inequality follows from $2B \geq p$. At the end of the day, the bound we have is of the form

$$\left|[-B, B[^{n+1} \cap \overline{E}_p\right| \leq 4^{n+1} \frac{B^{n+1}}{p^{n-1}}.$$

Let us now come back to the sum in (3.4), which we can split according to the two cases above:

$$\sum_{CB^2 > p > M} \frac{\left|\overline{E}_p \cap [-B, B[^{n+1}\right|}{(2B)^{n+1}} \leq \sum_{CB^2 > p > 2B} \frac{(2B)^2}{(2B)^{n+1}} + \sum_{2B \geq p > M} \frac{2^{n+1}}{p^{n-1}}. \tag{3.7}$$

Using the limit in B, the first sum goes to zero by the prime number theorem since $n \geq 3$. As B goes to infinity, the other sum becomes a converging series (again $n \geq 3$) starting at the index M. Letting M go to infinity, this too goes to zero. Hence we have shown that condition (3.1) holds, and the theorem follows. □

In degree 2, the above proof does not work: Indeed, it is easily seen that $\sum_p s_p$ diverges for $n = 2$, so by the first claim of Lemma 3.10, the proof we gave in degree greater or equal than 3 is doomed to fail in degree 2. However, we have a much simpler application of the lemma which shows that the density of shifted Eisenstein polynomials of degree 2 is indeed one, as Theorem 3.11 suggests.

Proposition 3.12. *The density of shifted Eisenstein polynomials of degree $n = 2$ is one.*

Proof. Let again $U_\infty = \emptyset$. We now apply Lemma 3.10 to a truncated sequence of sets. For this, let M be a positive integer and

$$U_p = \begin{cases} \overline{E}_p & \text{if } p \leq M \\ \emptyset & \text{if } p > M. \end{cases}$$

This truncated sequence now automatically satisfies condition (3.1), and we get the density

$$\underline{\mathbb{D}}(\overline{E}) \geq \mathbb{D}\Big(\bigcup_{p\leq M} \overline{E}_p \cap \mathbb{Z}^3\Big) = 1 - \prod_{p\leq M}\left(1 - \frac{(p-1)^2}{p^3}\right).$$

Letting M tend to infinity gives $\mathbb{D}(\overline{E}) = 1$, as the product diverges to zero. □

Remark 3.13. Even though the density of shifted Eisenstein polynomials of degree 2 is one, not all irreducible polynomials are Eisenstein for some shift (or even affine transformation): Take for example the polynomial $f(x) = x^2 + 8x - 16$, which is irreducible over $\mathbb{Z}$. Its discriminant is 2^7, so it could only be shifted Eisenstein with respect to 2. But neither $f(x)$ nor $f(x+1) = x^2 + 10x - 7$ is 2-Eisenstein.

3.3 Affine Eisenstein Polynomials

In [HS14, Section 7], the question was also raised about the density of polynomials that become Eisenstein after an arbitrary affine transformation, instead of only considering shifts. We can address this question as well, using the same methods as in Section 3.2.

Definition 3.14. For $f(x) \in R^{n+1}$ and $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in R^{2\times 2}$, we define the *affine transformation of f by A* as

$$f * A \coloneqq (cx+d)^n \cdot f\left(\frac{ax+b}{cx+d}\right).$$

It is easy to see that, when restricted to $\mathrm{GL}_2(R)$, this is a right group action.

Like in Section 3.2, we consider the set of polynomials with integer coefficients that become Eisenstein after some affine transformation.

Definition 3.15. Let $\widetilde{E} \subseteq \mathbb{Z}^{n+1}$ be the set of degree n polynomials which become Eisenstein of degree n after some affine transformation:

$$\widetilde{E} \coloneqq \{ f(x) \in \mathbb{Z}^{n+1} \mid f * A \in E \text{ for some } A \in \mathbb{Z}^{2\times 2}\}.$$

We call these polynomials *affine Eisenstein.*

It is easy to see that if both f and $f * A$ have degree n, and $f * A$ is irreducible, then so is f. Hence, an affine Eisenstein polynomial is irreducible.

Remark 3.16. The reader should notice that also in this case, we only consider affine Eisenstein polynomials of degree *exactly* n. Here, a further observation is required: It could happen that a degree n polynomial becomes Eisenstein of *lower* degree after some affine transformation. We don't consider such polynomials to be affine Eisenstein, since it can be seen that they are never irreducible. Likewise, a polynomial of degree less than n cannot become Eisenstein of degree n after an affine transformation, since any transformation that increases the degree introduces factors $cx + d$.

We again consider each prime separately by working over the p-adic integers.

Definition 3.17. Let $\tilde{E}_p \subseteq \mathbb{Z}_p^{n+1}$ be the set of degree n polynomials of $\mathbb{Z}_p[x]$ which become Eisenstein of degree n after some affine transformation $A \in \mathbb{Z}_p^{2\times 2}$:

$$\tilde{E}_p \coloneqq \{f(x) \in \mathbb{Z}_p^{n+1} \mid f * A \in E_p \text{ for some } A \in \mathbb{Z}_p^{2\times 2}\}.$$

We also call these polynomials affine Eisenstein, since it will always be clear from the context to which set we are referring.

In what follows, we compute the measure $\mu_p(\tilde{E}_p)$. For this, we need to write $\tilde{E}_p$ as a disjoint union of transformed copies of E_p as in Lemma 3.8. The following lemma is essential for this.

Lemma 3.18. *Assume $f(x) \in \mathbb{Z}_p^{n+1}$ is Eisenstein of degree $n \geq 2$, and let $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathbb{Z}_p^{2\times 2}$. Then, $f * A$ is Eisenstein of degree n if and only if $p \mid b$, $p \nmid a$, $p \nmid d$.*

Proof. If we write $f(x) = \sum_{i=0}^{n} \alpha_i x^i$ and $f * A = \sum_{l=0}^{n} \beta_l x^l$, then a simple calculation gives

$$\beta_l = \sum_{j=0}^{l} \sum_{s=l}^{n} \binom{n+j-s}{j} \binom{s-j}{l-j} \alpha_{s-j} d^{n-s} b^{s-l} a^{l-j} c^j. \tag{3.8}$$

Assume now that $f * A$ is Eisenstein, so $p \mid \beta_l$ for $0 \leq l \leq n-1$, $p^2 \nmid \beta_0$, $p \nmid \beta_n$. Consider first β_0. Reducing modulo p and using that $p \mid \alpha_i$ for $i < n$, we see that

$$\beta_0 \equiv \alpha_n b^n \pmod{p}.$$

Since $p \nmid \alpha_n$, we get that $p \mid b$. Knowing this, we reduce β_0 modulo p^2 and get

$$\beta_0 \equiv \alpha_0 d^n + \alpha_1 d^{n-1} b \equiv \alpha_0 d^n \pmod{p^2},$$

since $p^2 \mid \alpha_1 b$. From this, we see that $p^2 \nmid \beta_0$ if and only if $p \nmid d$. Finally, we reduce β_n modulo p and get

$$\beta_n \equiv \alpha_n a^n \pmod{p},$$

from which we conclude that $p \nmid a$.

Vice versa, if we assume that $p \mid b$, $p \nmid a$, $p \nmid d$, the same computations as above show that $p \nmid \beta_n$, $p \mid \beta_0$, $p^2 \nmid \beta_0$, and we easily see from (3.8) that $p \mid \beta_l$ for $0 < l < n$. Hence, $f * A$ is Eisenstein. □

We denote by $S = \left\{ \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathbb{Z}_p^{2\times 2} \;\middle|\; p \mid b,\ p \nmid a,\ p \nmid d \right\}$ the set of matrices from Lemma 3.18. This is a subgroup of $\mathrm{GL}_2(\mathbb{Z}_p)$. We can obtain the disjoint union decomposition of $\tilde{E}_p$ by considering the left cosets of S, but first, we need to deal with the noninvertible matrices. It turns out that they don't matter.

Lemma 3.19. *Let $n \geq 2$ and $f(x) \in \mathbb{Z}_p^{n+1}$. If $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathbb{Z}_p^{2\times 2}$ is* not *invertible, then $f * A$ is not Eisenstein of degree n.*

Proof. Assume for contradiction that $f * A$ is Eisenstein of degree n. We write again $f(x) = \sum_{i=0}^{n} \alpha_i x^i$ and $f * A = \sum_{l=0}^{n} \beta_l x^l$. We reduce modulo p:

$$\bar{A} = \begin{pmatrix} \bar{a} & \bar{b} \\ \bar{c} & \bar{d} \end{pmatrix} \in \mathbb{F}_p^{2\times 2}.$$

Since $\det \bar{A} = 0$, there are two cases: Either $\bar{c} = \bar{d} = 0$, or there is a $\lambda \in \mathbb{F}_p$ such that $\bar{a} = \lambda\bar{c}$ and $\bar{b} = \lambda\bar{d}$.

We consider the second case. Since $f * A$ is Eisenstein, we see that $\bar{f} * \bar{A} = \bar{\beta}_n x^n \in \mathbb{F}_p[x]$ with $\bar{\beta}_n \neq 0$. On the other hand,

$$\bar{f} * \bar{A} = (\bar{c}x + \bar{d})^n \cdot \bar{f}\left(\frac{\lambda\bar{c}x + \lambda\bar{d}}{\bar{c}x + \bar{d}}\right) = (\bar{c}x + \bar{d})^n \cdot \bar{f}(\lambda).$$

From this, we see that $\bar{f}(\lambda) \neq 0$, $\bar{c} \neq 0$ and $\bar{d} = 0$. This means that $p \mid d$ and $p \mid b$, from which it follows by (3.8) that $p^2 \mid \beta_0$. This contradicts the assumption that $f * A$ is Eisenstein.

The case $\bar{c} = \bar{d} = 0$ is similar. □

Hence, we only need to consider the action of $\mathrm{GL}_2(\mathbb{Z}_p)$ on $\mathbb{Z}_p^{n+1}$. According to Lemma 3.18, the action of elements of S does not change whether a polynomial is Eisenstein. Therefore, to see if a polynomial $f(x) \in \mathbb{Z}_p^{n+1}$ is affine Eisenstein, it is enough to check one representative of each left coset of $S \subset \mathrm{GL}_2(\mathbb{Z}_p)$. We can list these cosets explicitly.

Lemma 3.20. *The subgroup $S \subset \mathrm{GL}_2(\mathbb{Z}_p)$ has $p+1$ left cosets, which are the following:*

- $\begin{pmatrix} 1 & i \\ 0 & 1 \end{pmatrix} S$ *for $i \in \{0, \dots, p-1\}$ (corresponding to shifts), and*
- $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} S$ *(corresponding to the reciprocal).*

Proof. It is easy to see that these $p+1$ left cosets are distinct. We need to show that every $A = \left(\begin{smallmatrix} s & t \\ u & v \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{Z}_p)$ lies in one of them.

Consider first the case $p \mid v$. Then,

$$\begin{pmatrix} s & t \\ u & v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u & v \\ s & t \end{pmatrix},$$

with $\left(\begin{smallmatrix} u & v \\ s & t \end{smallmatrix}\right) \in S$. We have $p \nmid t$ and $p \nmid u$ because A is invertible.

If instead $p \nmid v$, let $i \equiv t/v \pmod{p}$, $i \in \{0, \dots, p-1\}$. Then,

$$\begin{pmatrix} s & t \\ u & v \end{pmatrix} = \begin{pmatrix} 1 & i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} s - iu & t - iv \\ u & v \end{pmatrix},$$

with $p \mid t - iv$ by choice of i, and $p \nmid s - iu$ since the matrix is invertible. □

Together, Lemmata 3.18 and 3.20 say that $f(x)$ is affine Eisenstein with respect to some A if and only if it is shifted Eisenstein with respect to some $i \in \{0, \dots, p-1\}$, or if its reciprocal $x^n \cdot f(1/x)$ is Eisenstein; and these possibilities are exclusive. In other words,

$$\tilde{E}_p = \text{reciprocal}(E_p) \sqcup \bigsqcup_{i=0}^{p-1} \sigma^{-i} E_p.$$

Since shifting and taking the reciprocal are linear maps with determinant ±1, they preserve the Haar measure, and we see that

$$\mu_p(\tilde{E}_p) = (p+1)\mu_p(E_p) = \frac{(p+1)(p-1)^2}{p^{n+2}}.$$

With this, we can now show the analogue of Theorem 3.11 for affine transformations.

Theorem 3.21. *Let $n \geq 3$. The density of affine Eisenstein polynomials of degree n is*

$$\mathbb{D}(\tilde{E}) = 1 - \prod_{p \text{ prime}} \left(1 - \frac{(p+1)(p-1)^2}{p^{n+2}}\right).$$

Proof. The proof is mostly the same as for Theorem 3.11. For the verification of condition (3.1), note that the case $2B < p$ is unchanged from the proof of Theorem 3.11, since the reciprocal polynomial cannot be p-Eisenstein for $p > B$. For the case $2B \geq p$, we simply get an additional term in the union (3.5), and so the estimate changes to $|\overline{L}_p| \leq p(p+1)$. However, this doesn't affect the convergence of the sum in (3.7). □

Remark 3.22. Clearly, the density of affine Eisenstein polynomials of degree $n = 2$ is one. After all, we are considering a superset of the shifted Eisenstein polynomials of Proposition 3.12.

3.4 Monte Carlo Simulations

As in [HS14, Section 6], we ran some Monte Carlo simulations to verify how near our results are to the actual probability of finding a shifted (or affine) Eisenstein polynomial among all the polynomials of a given height. For degrees $n = 3$ and 4, we tested 20 000 random polynomials of height at most 1 000 000. The results are shown in Tables 3.1 and 3.2. The first column contains the number of polynomials which were actually found by the Monte Carlo experiment, while the second column contains the expected number given by [HS13, Theorem 2] and Theorems 3.11 and 3.21. All the experiments seem to agree with our theoretical results.

The simulations were done using the Sage computer algebra system [Ste+14], and the code is available upon request.

Table 3.1: Simulations for degree $n = 3$.

	found	expected
irreducible	20 000	20 000
Eisenstein	1112	1112
shifted Eisenstein	3416	3353
affine Eisenstein	4360	4328

Table 3.2: Simulations for degree $n = 4$.

	found	expected
irreducible	20 000	20 000
Eisenstein	432	449
shifted Eisenstein	1096	1112
affine Eisenstein	1570	1547

Chapter 4

Irreducible Compositions of Polynomials

4.1 Introduction

Since irreducible polynomials play a fundamental role in applications and in the whole theory of finite fields (see for example [LN97; MP13; Rab+81; Bar+14; OS10; Ahm+12]), related questions have a long history (see for example [GHP99; GP97; Sho90; Jon12; JB12; And+14; Zur03]). In this chapter of my thesis, we specialize on irreducibility questions regarding compositions of polynomials. This kind of question has been addressed in the specific case of stable qudaratic polynomials, for which any repeated composition is irreducible, see for example in [Ahm09; GN10; Ahm+12; OS10; JB12; Jon12]. For analogous results related to additive polynomials, see [BM94a; BM94b].

In what follows, q will be an odd prime power, $\mathbb{F}_q[x]$ the univariate polynomial ring over the finite field $\mathbb{F}_q$ and $\mathrm{Irr}(\mathbb{F}_q[x])$ the set of irreducible polynomials in $\mathbb{F}_q[x]$.

The direct inspiration for this chapter is the paper by Jones and Boston [JB12], who give a criterion to decide whether a quadratic polynomial over $\mathbb{F}_q$ is stable, requiring only a finite amount of computation. The question then naturally arises whether this result can be extended to sets of multiple polynomials. We have the following motivating example. Let $q = 13$, and consider the quadratic polynomials $f = (x-5)^2 + 5$ and $g = (x-6)^2 + 5$. One can experimentally check that any possible composition of copies of f and g is irreducible. How can we prove that this is indeed the case?

More generally, let $\mathcal{S} \subset \mathbb{F}_q[x]$ be a set of monic quadratic polynomials, and let us denote by $C_{\mathcal{S}}$ the set of all compositions of copies of polynomials in $\mathcal{S}$. In other words, $C_{\mathcal{S}}$ is the monoid generated by $\mathcal{S}$ with the operation of composition. A couple of observations are now necessary:

- In principle, it is unclear whether a finite number of irreducibility checks will ensure that $C_{\mathcal{S}}$ is a subset of $\mathrm{Irr}(\mathbb{F}_q[x])$.

- It is unlikely that $C_{\mathcal{S}} \subseteq \mathrm{Irr}(\mathbb{F}_q[x])$ by chance, as the density of degree 2^n monic irreducible polynomials over $\mathbb{F}_q$ is roughly $1/2^n$. Thus, if $C_{\mathcal{S}}$ satisfies this property, one reasonably expects that there must be an algebraic reason for that.

In the first half of this chapter, which is based on our paper [FMS16], we address these issues by giving a necessary and sufficient condition for the monoid $C_{\mathcal{S}} \subset \mathbb{F}_q[x]$ to be contained in $\mathrm{Irr}(\mathbb{F}_q[x])$. This condition is algebraic and can be checked by performing only a finite amount of computation over $\mathbb{F}_q$, answering both points above.

Since it is quite rare that all compositions of a set of quadratic polynomials are irreducible, we ask in the second half whether we can describe exactly *which* compositions are irreducible in a general case. We accomplish this by applying the theory of finite automata, and show that the set of compositions which are irreducible correspond to a regular language, for which we can explicitly construct a finite state automaton. This second half is based on [FMS17].

We start off by giving some basic tools needed to establish our results in Section 4.2. Then, in Section 4.3, we describe the condition under which the compositional monoid consists only of irreducible polynomials and provide two nontrivial examples. In Section 4.3.1, we show the non-existence of such $C_{\mathcal{S}}$ whenever q is a prime congruent to 3 modulo 4 and there are two generating polynomials with no linear term (Proposition 4.11).

In Section 4.4, we give a short introduction to automata theory. Section 4.5 is concerned with the question of whether the compositional monoid generated by a set of quadratic polynomials is free. We then show in Section 4.6 that the irreducible elements of this monoid correspond to a regular language, and give a automaton that accepts it.

Finally, we show in Section 4.7 that the entire theory lifts to local fields under the assumption that the set of polynomials is finite and none of its elements have discriminant in the maximal ideal of the local field.

4.2 Capelli's Lemma

In this section, we describe the basic tools needed to establish the main results. We start with a well known result by Capelli, which gives a necessary and sufficient criterion to control the irreducibility of the composition of two polynomials.

Lemma 4.1 (Capelli's Lemma)**.** *Let K be a field and $f, g \in K[x]$ polynomials. Let $\beta \in \overline{K}$ be a root of g. Then, $g \circ f$ is irreducible over K if and only if g is irreducible over K and $f - \beta$ is irreducible over $K(\beta)$.*

See for example [FS96, Lemma 0.1] for a proof. We now use Capelli's Lemma to produce a simple ancillary result which will help us in what follows.

Lemma 4.2. *Let $g \in \mathbb{F}_q[x]$ be monic and irreducible of even degree, and let $f = (x - a_f)^2 - b_f \in \mathbb{F}_q[x]$. Then, $g \circ f$ is irreducible if and only if $g(-b_f)$ is not a square in $\mathbb{F}_q$.*

Proof. Let $d = \deg(g)$, and let $\beta \in \mathbb{F}_{q^d}$ be a root of g. According to Lemma 4.1, $g \circ f$ is irreducible over $\mathbb{F}_q$ if and only if $f - \beta$ is irreducible over $\mathbb{F}_{q^d}$. Writing $f - \beta = (x - a_f)^2 - (b_f + \beta)$, this is equivalent to $b_f + \beta$ not being a square in $\mathbb{F}_{q^d}$. Let $N : \mathbb{F}_{q^d} \longrightarrow \mathbb{F}_q$ be the norm map. If $\beta_1, \dots, \beta_d$ are the roots of g, we have

$$N(b_f + \beta) = \prod_{i=1}^{d}(b_f + \beta_i) = (-1)^d \prod_{i=1}^{d}((-b_f) - \beta_i) = g(-b_f),$$

since d is even. Now we can conclude, since $b_f + \beta$ is a nonsquare in $\mathbb{F}_{q^d}$ if and only if $N(b_f + \beta) = g(-b_f)$ is a nonsquare in $\mathbb{F}_q$. □

We are now ready to state one of the basic ingredients of the proof of the main theorem, which will allow us to consider irreducibility questions for compositions of degree two polynomials on a finite level.

Proposition 4.3. *Let $f_1, \dots, f_k \in \mathbb{F}_q[x]$ be monic polynomials of degree two. Write $f_i = (x - a_i)^2 - b_i$ for all i. Then, $f_1 \circ \cdots \circ f_k$ is irreducible if and only if all of the following are nonsquares in $\mathbb{F}_q$:*

- b_1
- $f_1(-b_2)$

  ⋮

- $(f_1 \circ \cdots \circ f_{k-1})(-b_k)$.

Proof. Clearly, f_1 is irreducible if and only if b_1 is a nonsquare. The rest follows by inductive application of Lemma 4.2. □

4.3 Irreducibility of the Entire Monoid

In this section, we give a criterion to determine whether the compositional monoid $C_{\mathcal{S}}$ generated by $\mathcal{S}$ consists only of irreducible polynomials. In order to state our result, we first need the following definition, which describes how to build a finite graph encoding only the useful (to our purposes) information contained in the generating set of the monoid.

Definition 4.4. Let q be an odd prime power, $\mathbb{F}_q$ the finite field of order q and $\mathcal{S}$ a subset of $\mathbb{F}_q[x]$. We denote by $G_{\mathcal{S}}$ the directed multigraph defined as follows:

- the set of nodes of $G_{\mathcal{S}}$ is $\mathbb{F}_q$;

- for any node $a \in \mathbb{F}_q$ and any polynomial $f \in \mathcal{S}$, there is a directed edge $a \longrightarrow f(a)$. We label that edge with f.

Before stating the next definition, we recall that for any monic polynomial f of degree 2 there exists a unique pair $(a_f, b_f) \in \mathbb{F}_q^2$ such that $f = (x - a_f)^2 - b_f$.

Definition 4.5. Let $\mathcal{S}$ be a subset of $\mathbb{F}_q[x]$ consisting of monic polynomials of degree 2. We call the set $D_{\mathcal{S}} \coloneqq \{-b_f \mid f \in \mathcal{S}\} \subseteq \mathbb{F}_q$, the *$\mathcal{S}$-distinguished set* of $\mathbb{F}_q$.

We are now ready to state and prove the main theorem.

Theorem 4.6. *Let $\mathcal{S}$ be a set of generators for a compositional monoid $C_{\mathcal{S}} \subseteq \mathbb{F}_q[x]$. Suppose that $\mathcal{S}$ consists of monic polynomials of degree 2. Then we have that $C_{\mathcal{S}} \subseteq \mathrm{Irr}(\mathbb{F}_q[x])$ if and only if no element of $-D_{\mathcal{S}} = \{b_f \mid f \in \mathcal{S}\} \subseteq \mathbb{F}_q$ is a square and in $G_{\mathcal{S}}$ there is no path of positive length from a node of $D_{\mathcal{S}}$ to a square of $\mathbb{F}_q$.*

Proof. Suppose there is a reducible composition $f_1 \circ \cdots \circ f_k \in C_{\mathcal{S}}$, for $k \geq 1$ and $f_i \in \mathcal{S}$ for all i. Then, by Proposition 4.3, either b_1 or some $(f_1 \circ \cdots \circ f_{l-1})(-b_l)$ with $2 \leq l \leq k$ is a square in $\mathbb{F}_q$. The former is an element of $-D_{\mathcal{S}}$, and the latter can be reached from the node $-b_l \in D_{\mathcal{S}}$ via the path with arrows labelled $(f_{l-1}, \ldots, f_1)$, so either case contradicts the condition in the theorem.

Vice versa, if some $b_f \in -D_{\mathcal{S}}$ is a square, then $f \in C_{\mathcal{S}}$ is reducible; and if there is a positive path with arrows labelled $(f_{k-1}, \ldots, f_1)$ from $-b_k \in D_{\mathcal{S}}$ to a square, then $f_1 \circ \cdots \circ f_k$ is reducible by Proposition 4.3. □

The reader should observe that this theorem is a generalization [JB12, Proposition 2.3], as the stability condition there is the same as the one given by our graph whenever the monoid we are considering has only one generator. It is useful to mention the following corollary, which is immediate.

Corollary 4.7. *Let $\mathcal{S}$ be a set of monic irreducible degree two polynomials and $C_{\mathcal{S}}$ defined as in Theorem 4.6. Then $C_{\mathcal{S}} \subseteq \mathrm{Irr}(\mathbb{F}_q[x])$ if and only if there is no path of positive length from a node of $D_{\mathcal{S}}$ to a square of $\mathbb{F}_q$.*

Proof. It is enough to observe that whenever $\mathcal{S} \subseteq \mathrm{Irr}(\mathbb{F}_q[x])$ then $-D_{\mathcal{S}}$ consists of nonsquares. □

In the following examples, we show two infinite families of such monoids generated by two polynomials when -1 is a square in $\mathbb{F}_q$. See [GU82, Theorem 1.8] for a proof that these families are infinite.

Example 4.8. Assume that -1 is a square in $\mathbb{F}_q$, and let $a \in \mathbb{F}_q$ such that both a and $b = a + 1$ are nonsquares. Define $f = (x - a)^2 + a$ and $g = (x - b)^2 + a$. In this situation, we have $D_{\mathcal{S}} = \{a\}$, and by assumption, $-a$, a and b are all nonsquares. Since $f(a) = g(b) = a$ and $f(b) = g(a) = b$, all paths in $G_{\mathcal{S}}$ starting from a end in a nonsquare, and the conditions of Theorem 4.6 are satisfied. Figure 4.1 shows the relevant part of the graph $G_{\mathcal{S}}$.

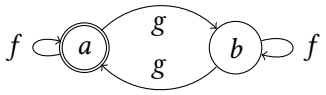

Figure 4.1: The nodes of $G_{\mathcal{S}}$ reachable from $D_{\mathcal{S}}$.

Example 4.9. Assume that -1 is a square in $\mathbb{F}_q$, and let $a \in \mathbb{F}_q$ such that a, $b = a+1$ and $c = a - 1$ are all nonsquares. Define $f = (x - a)^2 + a$ and $g = (x - a)^2 + c$. In this situation, we have $D_{\mathcal{S}} = \{a, c\}$, and by assumption, $\pm a$, $\pm b$ and $\pm c$ are all nonsquares. Again, Figure 4.2 shows the relevant part of the graph $G_{\mathcal{S}}$, and it is evident that the condition of Theorem 4.6 is satisfied.

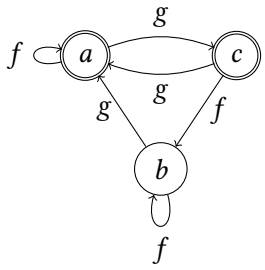

Figure 4.2: The nodes of $G_{\mathcal{S}}$ reachable from $D_{\mathcal{S}}$.

4.3.1 Non-Existence Results for $p \equiv 3 \bmod 4$

Whenever $q = p$ is a prime congruent to 3 modulo 4, we have the following non-existence results for polynomials without linear terms.

Proposition 4.10. *Let $p \equiv -1 \mod 8$ be a prime, and let $f = x^2 - b$ be a polynomial in $\mathbb{F}_p[x]$. Let C_f be the monoid generated by f. Then C_f contains a reducible polynomial, i.e. f is not stable.*

Proof. Assume for contradiction that $C_f \subset \mathrm{Irr}(\mathbb{F}_p[x])$. First note that if b is a square, then f is reducible, so we can assume that b is not a square, and thus $-b$ is a square. Consider the set of iterates $T = \{f(-b), f^2(-b), \ldots\} \subseteq \mathbb{F}_p$. By Corollary 4.7, C_f contains only irreducible polynomials if and only if T contains only nonsquares. So assume that this condition holds. Since T is finite, there exist $k < m \in \mathbb{N}_{>0}$ such that $f^m(-b) = f^k(-b)$. Choose k to be minimal. Now there are two cases: If $k > 1$, then there exist two distinct elements $u, v \in T$ such that $u^2 - b = v^2 - b$. Thus, $u = -v$, which implies that one between u and v is a square, a contradiction. If on the other hand $k = 1$, then we have $f^m(-b) = f(-b) = b^2 - b$, and so $f^{m-1}(-b)$ is either $-b$ or b. It can't be $-b$, since that is a square, so we must have $f^{m-1}(-b) = b \in T$. Setting $u = f^{m-2}(-b)$, we get that $u^2 - b = b$ and so $u^2 = 2b$, which is a contradiction because 2 is a square in $\mathbb{F}_p$ and consequently $2b$ is not. □

Proposition 4.11. *Let $p \equiv 3 \mod 4$ be a prime. Let $f = x^2 - b_f$ and $g = x^2 - b_g$ be polynomials in $\mathbb{F}_p[x]$ with b_f, b_g distinct nonsquares. Let $\mathcal{S} = \{f, g\}$ and let $C_\mathcal{S}$ be the monoid generated by $\mathcal{S}$. Then $C_\mathcal{S}$ contains a reducible polynomial.*

Proof. Let $G_\mathcal{S}$ be the graph attached to $\mathcal{S}$ as in Definition 4.4. Let $G'_\mathcal{S}$ be the induced subgraph consisting of all nodes of $G_\mathcal{S}$ that are reachable by some path of positive length starting from $-b_f$ or $-b_g$. That is, the edges of $G'_\mathcal{S}$ are just the edges of $G_\mathcal{S}$ starting and ending at a node in $G'_\mathcal{S}$. From now on, when we speak of nodes and edges, we will always be referring to nodes and edges in $G'_\mathcal{S}$. We call an edge from u to v an *f-edge* if it comes from the relation $f(u) = v$, while we call it a *g-edge* if it comes from $g(u) = v$. Since b_f and b_g are assumed nonsquare, we have by Corollary 4.7 that $C_\mathcal{S}$ contains a reducible polynomial if and only if at least one of the nodes of $G'_\mathcal{S}$ is a square. In the following, we assume for contradiction that $G'_\mathcal{S}$ consists only of nonsquares.

Let us observe the following: Suppose that there exists a node v of $G'_\mathcal{S}$ which is the target of two f-edges. By definition, this means that there exist two distinct nodes u, $u' \in G'_\mathcal{S}$ such that $u^2 - b_f = u'^2 - b_f = v$. This implies that $u' = -u$, and thus one between u and u' is a square, since -1 is not a square in $\mathbb{F}_p$. This contradicts our assumption. By symmetry, the same applies to g-edges.

By the argument above, we see that every node is the target of at most one f-edge and one g-edge, and by counting edges that it is indeed exactly one of each.

Now, consider the sum

$$\sum_{v \in G'_\mathcal{S}} (f(v) - g(v)).$$

On one hand, each node $u \in G'_\mathcal{S}$ appears exactly once as $f(v)$ and once as $g(v')$ for some v, $v' \in G'_\mathcal{S}$, so the sum is zero. On the other hand, it clearly holds that $f(v) - g(v) = b_g - b_f$ for all v. Letting n be the number of nodes in $G'_\mathcal{S}$, we get the equation

$$0 = n(b_g - b_f) \text{ in } \mathbb{F}_p.$$

Since $b_f \neq b_g$ by hypothesis, we must have $p \mid n$. This is impossible however, since $G'_\mathcal{S}$ is not empty and consists only of nonsquares, so $1 \leq n \leq \frac{p-1}{2}$. □

The fact that the polynomials of Proposition 4.11 don't have a linear term is of crucial importance. Let us see why by giving an explicit example of a monoid of irreducible polynomials in $\mathbb{F}_p[x]$ for which Proposition 4.11 does not apply (but $p \equiv 3 \mod 4$).

Example 4.12. Let us fix $p = 7$ and

$$f = (x-1)^2 - 5 = x^2 + 5x + 3 \in \mathbb{F}_7[x],$$
$$g = (x-4)^2 - 5 = x^2 + 6x + 4 \in \mathbb{F}_7[x].$$

The set $\mathcal{S} = \{f, g\}$ has distinguished set $D_\mathcal{S} = \{-5\}$ and graph as in Figure 4.3. Since 5 is not a square, and we only look at paths of positive length, the final claim follows by checking that 3 and -1 are not squares modulo 7.

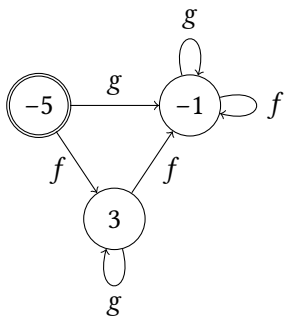

Figure 4.3: The nodes of G_S reachable from −5.

It is worth noting that this is one of only two examples we found with $p \equiv 3 \mod 4$, the other having $p = 103$.

4.4 Interlude on Automata Theory

We give a brief introduction to the basic definitions and results of automata theory, limiting ourselves to what we will use below. For a more extensive introduction, consult for example [HMU01].

We define an *alphabet* Σ as simply a finite set of symbols. A *word* over Σ is then a finite sequence of such symbols. We write Σ^* for the set of all words over Σ; so Σ^* is the free monoid generated by Σ. A *language* is then an arbitrary subset $\mathscr{L} \subseteq \Sigma^*$.

We now define a *nondeterministic finite state automaton*[1] (NFA) $\mathscr{M}$ over the alphabet Σ as a directed multigraph whose nodes are called *states* and whose arrows are called *transitions* and are labelled by symbols of Σ. Furthermore, a subset of states are designated as *start states*, and another as *accepting states*.

Consider now a word $w = a_1 \cdots a_t \in \Sigma^*$. We say that the automaton $\mathscr{M}$ *accepts* w if there is a path in $\mathscr{M}$ from a start state to an accepting state whose arrows are labelled in order a_1 to a_t. The *language accepted by* $\mathscr{M}$ consists of just those words it accepts. We say that a language is *regular* if there exists some NFA which accepts it. Regular languages have a particularly simple structure, and it is possible to completely describe one with what is called a regular expression.

A special case of an NFA is a *deterministic finite state automaton* (DFA). It has the additional restrictions that there exists exactly one start state, and for each pair of a state of $\mathscr{M}$ and a symbol in Σ, there exists at most one transition starting at that state and labelled by that symbol. The main benefit of a DFA is that it is easier to decide whether a particular word is accepted by it, since there is at most one path from the start state along any particular word.

[1] This is not the usual definition of an automaton in terms of a set of states and a transition function, but it is equivalent and serves our purposes better.

The following theorem is fundamental to automata theory, and is important for our construction below.

Theorem 4.13. *For any regular language, there exists a* deterministic *finite state automaton which accepts it.*

In other words, NFA are no more powerful than DFA. The proof of Theorem 4.13 involves the so called *subset construction*: From an NFA $\mathscr{M}$, we construct a DFA $\mathscr{M}_{\text{det}}$ as follows:

- The states of $\mathscr{M}_{\text{det}}$ are the subsets of states of $\mathscr{M}$.
- For each such state S of $\mathscr{M}_{\text{det}}$ and each symbol $a \in \Sigma$, we have a transition $S \xrightarrow{a} \{s \text{ state of } \mathscr{M} \mid \text{there is a transition } t \xrightarrow{a} s \text{ for some } t \in S\}$.
- The start state of $\mathscr{M}_{\text{det}}$ is the set of start states of $\mathscr{M}$.
- The accepting states of $\mathscr{M}_{\text{det}}$ are all states which contain an accepting state of $\mathscr{M}$.

One can now show that the automata $\mathscr{M}$ and $\mathscr{M}_{\text{det}}$ accept the same language.

Another way to look at regular languages is the following: Denote by $\cdot$ the concatenation operation of Σ^*. We define the *Kleene star* * as the operator which associates to a language $\mathscr{L}$ another language $\mathscr{L}^*$, which is the smallest submonoid of Σ^* containing $\mathscr{L}$. Then, a language is regular if and only if it is finite or can be expressed recursively starting from finite sets using the operations $\cup, \cdot, ^*$.

4.5 Freedom of the Compositional Monoid

Let $\mathbb{F}_q$ again be a finite field of odd characteristic and let $\mathcal{S} \subset \mathbb{F}_q[x]$ be a set of monic degree two polynomials. Recall that $C_{\mathcal{S}} \subseteq \mathbb{F}_q[x]$ is the compositional monoid generated by $\mathcal{S}$. We consider the set $\mathcal{S}$ to be an alphabet, and $\mathcal{S}^*$ is the set of words over the alphabet $\mathcal{S}$, i.e. the free monoid generated by the symbols in $\mathcal{S}$. A word $f_1 \cdots f_k \in \mathcal{S}^*$ corresponds to the composition $f_1 \circ \cdots \circ f_k \in \mathbb{F}_q[x]$ via the natural surjective morphism of monoids $\pi : \mathcal{S}^* \to C_{\mathcal{S}}$. The empty word naturally corresponds to x. Let $\mathcal{I} \subset \mathcal{S}^*$ be the language of words whose corresponding compositions are irreducible. Our goal is to show that $\mathcal{I}$ is a regular language by providing an automaton that accepts it.

As before, we write each polynomial $f \in \mathcal{S}$ as $f = (x - a_f)^2 - b_f$ for some a_f, $b_f \in \mathbb{F}_q$, and the set $D_{\mathcal{S}} \coloneqq \{-b_f \mid f \in \mathcal{S}\} \subseteq \mathbb{F}_q$ is called the $\mathcal{S}$-distinguished set of $\mathbb{F}_q$.

We include in this section some elementary facts concerning the freedom of the monoid $C_{\mathcal{S}}$ generated by a finite set of irreducible degree two polynomials. These results will be needed in Section 4.6. For $b \in D_{\mathcal{S}}$, we define A_b as the subset

of all a in $\mathbb{F}_q$ such that there exists $f \in \mathcal{S}$ with $f = (x-a)^2 - b$. For any of the A_b, we define the difference set

$$A_b - A_b = \{a - a' \mid a, a' \in A_b\}.$$

We can define a relation $\sim$ on $\mathcal{S}^*$ by setting $u \sim w$ if there exists $l \in \bigcup_{b \in D_\mathcal{S}} (A_b - A_b)$ for which $\pi(u) + l = \pi(w)$. This relation is symmetric and reflexive but not transitive, unless $\bigcup_{b \in D_\mathcal{S}} (A_b - A_b)$ is an additive subgroup of $\mathbb{F}_q$.

In this section, we provide a computable condition to establish whether $C_\mathcal{S}$ is a free monoid, which will be needed later on.

Proposition 4.14. *Let u, v be words of $\mathcal{S}^*$ of equal length $n \geq 1$. Let u', v' be the $(n-1)$-suffixes of u and v respectively. Then*

(i) *$\pi(u) = \pi(v)$ implies $u' \sim v'$,*

(ii) *$u \sim v$ if and only if $\pi(fu) = \pi(gv)$ for some $f, g \in \mathcal{S}$.*

Proof. Let us first prove (i). Suppose $\pi(u) = \pi(v)$ and let us write

$$\pi(u) = (h_1 - a)^2 - b = (h_2 - a')^2 - b' = \pi(v)$$

for $h_1 = \pi(u')$, $h_2 = \pi(v')$. Then we have $(h_1 - a - h_2 + a')(h_1 + h_2 - a - a') = b - b'$. Since $h_1 + h_2 - a - a'$ has positive degree, this forces $b = b'$ and $h_1 - a - h_2 + a' = 0$. Now it is clear that $a, a' \in A_b$, which implies $a' - a \in A_b - A_b$, and then $u' \sim v'$.

Let us now prove (ii). If $u \sim v$, by definition we have $\pi(u) - a = \pi(v) - a'$ for $a - a' \in A_b - A_b$ for some b. Now, by squaring and subtracting b on both sides of the equality we get $f(\pi(u)) = g(\pi(v))$ for some $f, g \in \mathcal{S}$, and hence $\pi(fu) = \pi(gv)$. Vice versa, if there exists $f, g \in \mathcal{S}$ such that $\pi(fu) = \pi(gv)$, then (i) applies. □

Lemma 4.15. *Let u, v be words of $\mathcal{S}^*$ of equal length $n \geq 1$. If $|D_\mathcal{S}| = |\mathcal{S}|$ or $|D_\mathcal{S}| = 1$, then we have that $u \sim v$ if and only if $\pi(u) = \pi(v)$.*

Proof. One direction is trivial: If $\pi(u) = \pi(v)$ then $u \sim v$. For the other direction, we look at the two cases separately.

In the case $|D_\mathcal{S}| = |\mathcal{S}|$, it follows from $u \sim v$ that $\pi(u) - \pi(v) = c \in A_b - A_b$ for some $b \in D_\mathcal{S}$. However, A_b consists of only one element, so $c = 0$.

For the case $|D_\mathcal{S}| = 1$, assume that $u \sim v$, so $\pi(u) = \pi(v) + c$ for some $c \in \mathbb{F}_q$. Let u', v' be the $(n-1)$-suffixes of u and v. Then, since $|D_\mathcal{S}| = 1$, we have that $(\pi(u') - a_1)^2 - (\pi(v') - a_2)^2 = c$ for some $a_1, a_2 \in \mathbb{F}_q$. As c is constant, this forces $(\pi(u') - a_1) - (\pi(v') - a_2) = 0$, which in turn forces $c = 0$ and hence $\pi(u) = \pi(v)$. □

The following proposition shows that the freedom of the monoid is ensured whenever $D_\mathcal{S}$ is either maximal or minimal.

Proposition 4.16. *If $|D_\mathcal{S}| = |\mathcal{S}|$ or $|D_\mathcal{S}| = 1$, then $C_\mathcal{S} \cong \mathcal{S}^*$.*

Proof. Clearly, a polynomial of degree two in $C_{\mathcal{S}}$ cannot have two distinct writings in terms of compositions. Let F be a polynomial in $C_{\mathcal{S}}$ of minimal degree with two different writings, i.e. such that $F = \pi(fu) = \pi(gv)$ for $f, g \in \mathcal{S}$ and $u, v \in \mathcal{S}^*$ of positive length. From $\pi(fu) = \pi(gv)$, one deduces by Proposition 4.14 that $u \sim v$. Lemma 4.15 now gives $\pi(u) = \pi(v)$, which implies $u = v$ by the minimality of F. □

Corollary 4.17. *If $|\mathcal{S}| = 2$ then $C_{\mathcal{S}} \cong \mathcal{S}^*$.*

Proof. Immediate by observing that $|D_{\mathcal{S}}| = 1$ or $|D_{\mathcal{S}}| = |\mathcal{S}| = 2$. □

4.6 An Automaton for Irreducible Compositions

We are now ready to prove that the language $\mathcal{I}$ of words over $\mathcal{S}$ which correspond to irreducible polynomials is regular. We do this by constructing an automaton accepting it. First, however, we define a different DFA $\mathcal{N} = \mathcal{N}(\mathcal{S})$ which checks the conditions of Proposition 4.3.

Definition 4.18. The states of the automaton $\mathcal{N}$ are given by the following:

- A special start state $\mathfrak{I}$. It is accepting.
- For each $a \in \mathbb{F}_q$, we have a distinguished state $[a]$. It is accepting if $-a$ is a nonsquare.
- For each $a \in \mathbb{F}_q$, we have a state $\{a\}$. It is accepting if a is a nonsquare.

The transitions are as follows:

- For each $f \in \mathcal{S}$, we have a transition $\mathfrak{I} \xrightarrow{f} [-b_f]$.
- For each $f \in \mathcal{S}$ and each $a \in \mathbb{F}_q$, we have a transition $[a] \xrightarrow{f} \{f(a)\}$.
- For each $f \in \mathcal{S}$ and each $a \in \mathbb{F}_q$, we have a transition $\{a\} \xrightarrow{f} \{f(a)\}$.

Remark 4.19. The reason we distinguish between the states $\{a\}$ and $[a]$ is that they may be accepting at different times: $\{a\}$ accepts if a is nonsquare, $[a]$ if $-a$ is nonsquare. In the case that -1 is a square in $\mathbb{F}_q$, the two are equivalent and we can identify the two types of states.

Theorem 4.20. *The language $\mathcal{I}$ of irreducible compositions is regular.*

Proof. Let $\mathcal{L}$ be the regular language over the alphabet $\mathcal{S}$ that is accepted by the automaton $\mathcal{N}$ reading from *right to left*. It is easy to see that a single letter f is in $\mathcal{L}$ if and only if b_f is nonsquare. Furthermore, a word $f_1 \cdots f_k$, $k \geq 2$, is in $\mathcal{L}$ if and only if $(f_1 \circ \ldots \circ f_{k-1})(-b_{f_k})$ is nonsquare. By Proposition 4.3, it follows that the word $f_1 \cdots f_k$ corresponds to an irreducible polynomial if and only if each prefix

$f_1 \cdots f_l$, $l \leq k$, lies in $\mathscr{L}$. In other words, $\mathscr{I}$ is the language of all words whose every prefix is in $\mathscr{L}$.

Let $\mathscr{N}_L$ now be a deterministic automaton that accepts $\mathscr{L}$ reading from the *left*. We obtain the automaton $\mathscr{M}$ accepting $\mathscr{I}$ from this by simply removing all non-accepting states; and it follows that $\mathscr{I}$ is regular. □

In order to actually construct the automaton $\mathscr{M}$, we note that we can obtain a *nondeterministic* automaton accepting $\mathscr{L}$ from the left by reversing the direction of all transitions and swapping start and accept states of $\mathscr{N}$. From this, we obtain $\mathscr{N}_L$ by the subset construction. We now provide an example to see the theorem in action.

Example 4.21. Consider the case $q = 5$ and $\mathcal{S} = \{f, g\}$ with $f = x^2 - 2$ and $g = (x-1)^2 - 3$.

We first construct the automaton $\mathscr{N}$. Since $p \equiv 1 \mod 4$, we can identify the nodes $[a]$ and $\{a\}$. Note that we have removed the node $\{0\}$ since it is not reachable from $\mathfrak{I}$. The result is seen in Figure 4.4.

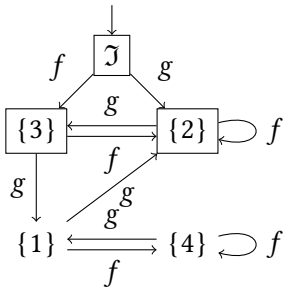

Figure 4.4: The automaton $\mathscr{N}$ for Example 4.21. Boxed states are accepting.

After performing the transformation described in the proof of Theorem 4.20 and cutting out all unreachable states, we end up with the simple deterministic automaton $\mathscr{M}$ in Figure 4.5. This shows that the irreducible compositions of f and g are precisely those of the form f^n, $f^n g$, $f^n g^2$ and $f^n g^2 f$ for $n \geq 0$. Here, multiplication means composition.

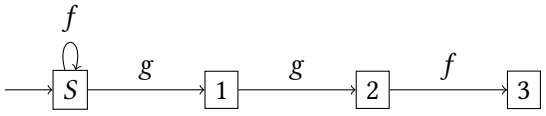

Figure 4.5: The automaton $\mathscr{M}$ accepting $\mathscr{I}$ for Example 4.21. All states are accepting.

Using the machinery we developed in the rest of this chapter, we describe an infinite set of primes of $\mathbb{F}_q[x]$ having a finite regular structure.

Theorem 4.22. *Let $\mathbb{F}_q$ be a finite field of characteristic different from 2. The set of monic irreducible polynomials having coefficients in $\mathbb{F}_q$ which can be written as a nonempty composition of degree 2 polynomials can be partitioned into a finite disjoint union $\bigsqcup_{a\in\mathbb{F}_q} \mathscr{L}_a$ in such a way that each $\mathscr{L}_a$ is in natural bijection with the words of a regular language $\mathscr{L}$, which is independent of a. In particular, the set of such irreducible polynomials has a finite regular expression in terms of the elementary operations $\cup, \cdot, ^*$.*

Proof. Let D be the set of monic irreducible polynomials in $\mathbb{F}_q[x]$ that can be written as nonempty composition of degree 2 polynomials. Let $\mathcal{S} = \{x^2 - b \mid b \in \mathbb{F}_q\}$. By Proposition 4.16, $C_{\mathcal{S}}$ is isomorphic to $\mathcal{S}^*$, so it is naturally embedded in $\mathbb{F}_q[x]$. Apply now Theorem 4.20 to obtain the regular language of irreducible polynomials $\mathscr{I}$ generated by $\mathcal{S}$, and let $\mathscr{L} = \mathscr{I} \setminus \{x\}$. Let $\psi_a : \mathscr{L} \longrightarrow \mathbb{F}_q[x]$ be the shift map defined by $f(x) \mapsto f(x + a)$. Let $\mathscr{L}_a = \psi_a(\mathscr{L})$. It is easy to observe that for any polynomial $f \in D$, there exists $a \in \mathbb{F}_q$ such that $f(x - a)$ can be written as an element of $C_{\mathcal{S}}$. This shows that

$$D = \bigcup_{a\in\mathbb{F}_q} \mathscr{L}_a.$$

It remains to show that $\mathscr{L}_a \cap \mathscr{L}_b = \emptyset$ if $a \neq b$, the final result will follow immediately. We argue by induction on the length of the words in $\mathscr{L}$ (i.e. the degree of the polynomials). Let $a, b \in \mathbb{F}_q$ with $a \neq b$ such that there exist two words $v, w \in \mathscr{L}$ of minimal length l such that $\psi_a(v) = \psi_b(w)$. If $l = 1$, this is clearly impossible, so let us assume $l > 1$. We can write $f(v'(x + a)) = g(w'(x + b))$ for some $f, g \in \mathcal{S}$ and v', w' suffixes of v and w respectively. Therefore, for some k, $j \in \mathbb{F}_q$ we have

$$\begin{aligned} v'(x + a)^2 - w'(x + b)^2 &= (v'(x + a) - w'(x + b))(v'(x + a) + w'(x + b)) \\ &= k - j. \end{aligned}$$

Since v', w' are monic and the characteristic of $\mathbb{F}_q$ is different from 2, then the degree of the polynomial $(v'(x + a) + w'(x + b))$ is greater than or equal to 2. This forces both $k = j$ and $(v'(x + a) - w'(x + b)) = 0$, which contradicts the minimality of l. □

Example 4.23. For an example demonstrating Theorem 4.22, take $q = 3$ and $\mathcal{S} = \{f, g, h\}$ with $f = x^2$, $g = x^2 - 1$, $h = x^2 - 2$. From Proposition 4.16, $C_{\mathcal{S}}$ is free and isomorphic to $\mathcal{S}^*$. Applying the construction, we get the automaton shown in Figure 4.6. We see that the irreducible polynomials in $C_{\mathcal{S}}$ are exactly x, h, hgf^n for $n \geq 0$, and h^2k for $k \in C_{\mathcal{S}}$ arbitrary (possibly the identity). Applying Theorem 4.22, it follows that the set of irreducible polynomials in $\mathbb{F}_3[x]$ that can be written as a nonempty composition of degree 2 polynomials is precisely

$$\bigcup_{a\in\mathbb{F}_3} (\{h(x + a)\} \cup \{hgf^n(x + a) \mid n \geq 0\} \cup \{h^2k(x + a) \mid k \in C_{\mathcal{S}}\}).$$

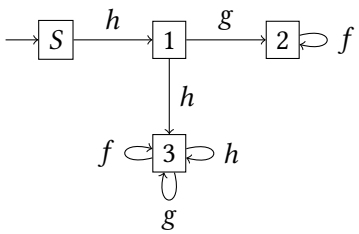

Figure 4.6: The automaton $\mathscr{M}$ accepting $\mathscr{I}$ for Example 4.23. All states are accepting.

4.7 Irreducible Compositions over Local Fields

In this final section, we will show how the results of the previous sections can be lifted, under some additional hypothesis, to polynomials over local fields. Let K be a non-archimedean local field with finite residue field $\mathbb{F}_q$ of odd characteristic. Let $\mathcal{O}_K$ be its ring of integers and ϖ be a uniformizer. We will denote by $\tilde{\cdot}$ the reduction map $\mathcal{O}_K[x] \longrightarrow \mathbb{F}_q[x]$. Let us start by recalling the following lemma, which we state in a weaker form, sufficient for our purposes.

Lemma 4.24. *Let L be any field and $f, g \in L[x]$ be monic polynomials of degree d_f, d_g, respectively. Suppose that $f = (x - a_f)^2 - b_f$ for some $a_f, b_f \in L$. Then we have:*

$$\operatorname{disc}(g \circ f) = \pm \operatorname{disc}(g)^{d_f} \cdot d_f^{d_g d_f} \cdot g(-b_f).$$

Proof. See [Jon08, Lemma 2.6]. □

Theorem 4.25. *Let $f_1, \dots, f_k \in \mathcal{O}_K[x]$ be monic polynomials of degree 2 such that $\varpi \nmid \operatorname{disc}(f_1)$. Then $f_1 \circ \cdots \circ f_k$ is irreducible in $K[x]$ if and only if $\tilde{f}_1 \circ \cdots \circ \tilde{f}_k$ is irreducible in $\mathbb{F}_q[x]$.*

Proof. One direction is obvious, so let us assume that $f_1 \circ \cdots \circ f_k$ is irreducible. For every $i = 1, \dots, k$, let $f_i = (x - a_i)^2 - b_i$ for some $a_i, b_i \in \mathcal{O}_K$. By Proposition 4.3, we need to show that the following elements are not squares:

- $c_1 \coloneqq \tilde{b}_1$
- $c_2 \coloneqq \tilde{f}_1(-\tilde{b}_2)$

  ⋮

- $c_k \coloneqq (\tilde{f}_1 \circ \cdots \circ \tilde{f}_{k-1})(-\tilde{b}_k)$.

First, suppose that $c_t = 0$ for some $t \in \{1, \dots, k\}$. This implies that $\tilde{f}_1$ has a root, and since by hypothesis the discriminant of $\tilde{f}_1$ is nonzero, by Hensel's Lemma we can lift such a root to a root of f_1. But then $f_1 \circ \cdots \circ f_k$ is clearly reducible, which is a contradiction. Thus we can assume that $c_i \neq 0$ for all $i \in \{1, \dots, k\}$. Now let

$t \in \{1, \dots, k\}$ be such that c_t is a nonzero square. By Proposition 4.3, this implies that $\tilde{f}_1 \circ \cdots \circ \tilde{f}_t$ is reducible. On the other hand, applying Lemma 4.24 recursively and using the definition of the c_i we get that

$$\operatorname{disc}(\tilde{f}_1 \circ \cdots \circ \tilde{f}_t) = u \cdot \prod_{i=1}^{t} c_i^{2^{t-i}} \neq 0,$$

where u is an appropriate power of 2 (up to sign). This proves that $\varpi \nmid \operatorname{disc}(f_1 \circ \cdots \circ f_t)$ and since $f_1 \circ \cdots \circ f_t$ is irreducible by hypothesis, it defines an unramified extension of K. It follows that $\tilde{f}_1 \circ \cdots \circ \tilde{f}_t$ is irreducible (see for example [Cas86, Chapter 7]), giving a contradiction. □

It is clear that the hypothesis that $\varpi \nmid \operatorname{disc}(f_1)$ is necessary for the claim to hold, since for example $x^2 - \varpi$ is irreducible in $K[x]$, while its reduction is reducible in $\mathbb{F}_q[x]$.

Given a finite set $\mathcal{S} \subseteq \mathcal{O}_K[x]$ of monic polynomials of degree two with unitary discriminant, Theorem 4.25 shows that irreducible compositions of the elements of $\mathcal{S}$ correspond bijectively to irreducible compositions of the elements of $\widetilde{\mathcal{S}} \subseteq \mathbb{F}_q[x]$. Therefore, if we consider $\mathcal{S}$ as an alphabet and $\mathcal{I}$ is the language of irreducible compositions of the elements of $\mathcal{S}$, we deduce immediately the following corollary.

Corollary 4.26. *The language $\mathcal{I}$ is regular.*

Proof. It is enough to apply Theorem 4.20 to the language of irreducible compositions of the elements of $\widetilde{\mathcal{S}}$. □

The above corollary essentially states that the theory we developed in the rest of the paper lifts entirely to local fields, at least in the case in which the elements in $\mathcal{S}$ have unitary discriminant. It would be interesting to understand what happens when this condition is not satisfied.

4.8 Open Questions

One of the natural questions arising from our investigation is whether our results can be generalised to higher degree polynomials. In fact any lift of such results to polynomials of degree three or more would be of great interest, as most of the literature is focused only on degree two, and in particular the necessary and sufficient criterion by Boston and Jones [JB12] (and the subsequent results on the subject such as [HM16; Ahm+12; Ahm09; GN10; FMS16]) only exists in degree two. In the context of local fields, another interesting issue arising from Theorem 4.25 of this paper is the following: How can one include singular polynomials in the generating set $\mathcal{S}$? In fact, the condition on the discriminant seems to be essential.

Chapter 5

Group Key Exchange

5.1 Introduction

Traditional cryptographic tools for key exchange may not be useful when the communication process is carried out in a group of nodes or users. There exist several approaches for group key management, which may be divided into three main classes [RH03]:

- *centralized* protocols, where a single entity is in charge of controlling the whole group, minimizing storage requirements and computational power on both the client and server side as well as communication overheads,
- *decentralized*, where a large group is divided into subgroups in order to avoid concentrating the workload in a single point,
- *distributed*, where key generation is carried out in a distributed and collaborative way.

This last class of approaches has become particularly important since the emergence of ad hoc networks, where a set of nodes, possibly consisting of light and mobile devices, create, operate and manage a network, which is therefore solely dependent on the cooperative and trusting nature of the nodes. Moreover the limited capacity of the involved devices imposes both key storage and computational requirements. Such a network is commonly created to meet an immediate demand and specific goal, and nodes are continuously joining or leaving it. Thus, group key management based on distributed and collaborative schemes has proved to be of great interest (see for instance [MDM07; LLY06] and their references).

Some efficient solutions were introduced by Burmester and Desmedt in [BD94] and [BD05], and by Steiner et al. in [STW96] and [STW00]. These protocols naturally extend the classical Diffie-Hellman protocol [DH76]. Both solutions were shown to be secure against a passive adversary if the Diffie-Hellman problem is intractable. The approach by Burmester and Desmedt is very efficient in the initial key agreement, using only two rounds. However, if the key has to be refreshed

or the group of users changes, the entire protocol needs to be rerun. On the other hand, the protocols by Steiner et al. require more rounds initially, but feature rekeying procedures that are more efficient than rerunning the protocol.

In [MMR07], the authors generalize the classical Diffie-Hellman key exchange to an arbitrary semigroup action $\Phi : G \times S \longrightarrow S$ of an abelian semigroup G on a set S. To simplify notation, we will write $g \cdot s \coloneqq \Phi(g, s)$ for $g \in G$ and $s \in S$ throughout.

Protocol 5.1 (Semigroup Diffie-Hellman Key Exchange)**.** Let S be a finite set, G an abelian semigroup, and $\Phi : G \times S \longrightarrow S$ a G-action on S. The semigroup Diffie-Hellman key exchange in (G, S, Φ) is the following protocol:

(1) Alice and Bob publicly agree on an element $s \in S$.

(2) Alice chooses $a \in G$ and computes $a \cdot s$. Alice's private key is a, her public key is $a \cdot s$.

(3) Bob chooses $b \in G$ and computes $b \cdot s$. Bob's private key is b, his public key is $b \cdot s$.

(4) Their common secret key is then

$$a \cdot (b \cdot s) = ab \cdot s = ba \cdot s = b \cdot (a \cdot s).$$

In the original Diffie-Hellman proposal, if an adversary is able to solve the so-called Discrete Logarithm Problem (DLP), then she is able to break the Diffie-Hellman key exchange. In this setting we can analogously consider the following more general problem:

Problem 5.2 (Semigroup Action Problem, SAP)**.** Given a semigroup G acting on a set S and elements $x, y \in S$, find $g \in G$ such that $g \cdot x = y$.

It is clear that if an adversary finds a $g \in G$ such that $g \cdot s = a \cdot s$, then she can find the shared secret by computing $g \cdot (b \cdot s) = gb \cdot s = bg \cdot s = b \cdot (a \cdot s)$.

We can say that the security of the preceding protocol is equivalent to the following problem.

Problem 5.3 (Diffie-Hellman Semigroup Action Problem, DHSAP)**.** Given a finite abelian semigroup G acting on a finite set S and elements $x, y, z \in S$ with $y = g \cdot x$ and $z = h \cdot x$ for some $g, h \in G$, find $gh \cdot x$.

Although, as noted above, solving the SAP implies solving the DHSAP, we do not know if both problems are (in general) equivalent, just like in the traditional setting of Diffie-Hellman, where however some equivalence results for particular scenarios are known [MW99].

Motivated by the above, our idea is now to define extensions of the semigroup Diffie-Hellman key exchange protocol to n users, by first generalizing those introduced in [STW96] and [STW00], and then considering other settings, which can

feature more favorable characteristics compared to the original protocol. Since the capability of devices is often limited and authentication processes may be difficult to implement in a distributed network, we focus our attention on confidentiality under passive attacks. As in [MMR07], some nonstandard settings are introduced as more general examples, although the hardness of the SAP there may not be proven yet, so the security of the protocols in those cases is conditional on that.

The first half of this chapter is based on [Lóp+15], and is structured as follows: In Section 5.2, we consider a suite of three protocols for group key management based on one-sided actions. While these naturally extend the results of [STW96] and [STW00], we consider different settings for a general semigroup action. Section 5.3 considers the security of the preceding protocols against passive attacks, including forward and backward secrecy. Then, in Section 5.4, we introduce two protocols based on linear actions, i.e. semigroup actions on other groups satisfying a certain distributivity property. We give two different group key protocols in this setting, one of which runs very efficiently in only two rounds, independently of the number of members in the communicating group.

After this, in Section 5.5, we describe two more protocols based on linear actions, which are variants of the second one above and were published in [Lóp+16]. Section 5.6, which is based on [Sch+16], describes an active attack on the GSAP-3 protocol of [STW96; STW00] and our variant thereof. This attack is in some ways more powerful than a traditional man-in-the-middle attack. Finally, in Section 5.7, we recount a similar attack on the protocol of Burmester and Desmedt [BD94; BD05]. This attack was published in [Bao+16].

5.2 Group Key Communication based on One-Sided Actions

In this section we consider three different extensions of the semigroup Diffie-Hellman key exchange with different computing requirements and communication overheads, but with possible applications in different cases. They are natural extensions of [STW96] and [STW00]. For completeness we report proofs in appendix to show soundness of the schemes.

5.2.1 A Sequential Key Agreement

The first approach to extend the key exchange protocol consists of a sequence of messages, built using pieces of private information, along a chain of users and an analogous second sequence of messages in the opposite way. Therefore every user will send and receive two messages except for the user that initiates the communication and the last user receiving the sequence of messages.

We will consider a group of n users, $\mathcal{U}_1, \dots, \mathcal{U}_n$, who would like to share a secret element of a finite set S, and G will denote a finite abelian semigroup acting

on S.

The protocol is defined by the following steps.

Protocol 5.4 (GSAP-1)**.** Users agree on an element s in a finite set S, a finite abelian semigroup G, and a G-action on S given by Φ. For every $i = 1, \dots, n$, the user $\mathcal{U}_i$ holds a private element $g_i \in G$.

(1) For $i = 1, \dots, n-1$, user $\mathcal{U}_i$ sends to user $\mathcal{U}_{i+1}$ the message

$$\{C_1, \dots, C_i\} = \left\{ g_1 \cdot s,\ g_2 g_1 \cdot s,\ \dots,\ \left(\prod_{j=1}^{i} g_j \right) \cdot s \right\}.$$

(2) User $\mathcal{U}_n$ computes $g_n \cdot C_{n-1}$.

(3) For $k = n, \dots, 2$, user $\mathcal{U}_k$ sends to user $\mathcal{U}_{k-1}$ the message $\{f_1^k, \dots, f_{k-1}^k\}$, where $f_j^k = g_k \cdot f_j^{k+1}$ for $2 \le k \le n-1$ and $f_j^n = g_n \cdot C_{j-1}$ for $j = 1, \dots, n-1$, with $C_0 = s$.

(4) User $\mathcal{U}_k$ computes $g_k \cdot f_k^{k+1}$.

5.2.2 A Key Agreement in Broadcast

The following protocol presents a lower communication overhead than GSAP-1. The idea is again to get a first sequence of messages from user $\mathcal{U}_1$ to user $\mathcal{U}_n$, but now $\mathcal{U}_n$ will broadcast a message that allows the rest of the users to recover the common key.

Protocol 5.5 (GSAP-2)**.** Users agree on an element s in a finite set S, a finite abelian semigroup G, and a G-action Φ on S. For every $i = 1, \dots, n$, the user $\mathcal{U}_i$ holds a private element $g_i \in G$.

(1) For $i = 1, \dots, n-1$, user $\mathcal{U}_i$ sends to user $\mathcal{U}_{i+1}$ the message

$$\{C_{i-1}^{i-1}, C_1^i, \dots, C_i^i\},$$

where $C_0^0 = s$ and $C_1^1 = g_1 \cdot s$, and for $i \ge 2$, $C_1^i = g_i \cdot C_{i-2}^{i-2}$ and $C_j^i = g_i \cdot C_{j-1}^{i-1}$ (with $j = 2, \dots, i$).

(2) User $\mathcal{U}_n$ computes $g_n \cdot C_{n-1}^{n-1}$.

(3) User $\mathcal{U}_n$ broadcasts $\{f_1^n, \dots, f_n^n\}$, where $f_i^n = g_n \cdot C_{n-1-i}^{n-1}$ for $i = 1, \dots, n-2$ and $f_{n-1}^n = g_n \cdot C_{n-2}^{n-2}$ and $f_n^n = C_{n-1}^{n-1}$.

(4) User $\mathcal{U}_i$ computes $g_i \cdot f_i^n$.

Remark 5.6. It can be observed that the element f_n^n contained in the broadcast message in step (3) of Protocol 5.5, is not needed by any of the users $\mathcal{U}_i$, $i = 1, \dots, n-1$, to recover the shared key. However, the distribution of this value is required for future rekeying operations, as we will later show.

5.2.3 Examples

(a) The two previous protocols are extensions of those introduced in [STW96] and [STW00] for the action of the multiplicative semigroup $\mathbb{N}^*$ on a cyclic group S of order q generated by g, given by $x \cdot s = s^x$. It was pointed out that the first protocol presents excessive communication overheads due to the number of rounds and messages to be sent. Because of this, only the second one, referred to as IKA.1 in [STW00], was recommended. However, the first protocol could be interesting on its own when applied to a sensor network whose communications need to be secure and where it should be assessed whether every node is working properly. After user $\mathcal{U}_n$ receives the message in step (1), the absence of any of the messages (excepting the last one) in the descending chain of rounds would alert that the corresponding sender node is not working or the communication was interrupted.

(b) In particular, consider a finite field $\mathbb{F}_q$ and an element g of prime order. The semigroup $\mathbb{N}^*$ acts on the subgroup $\langle g \rangle \subset \mathbb{F}_q^*$ by $x \cdot s = s^x$ for $x \in \mathbb{N}^*$, $s \in \langle g \rangle$.

(c) Let ε be the set of points in an elliptic curve. Then the action $\Phi : \mathbb{N}^* \times \varepsilon \longrightarrow \varepsilon$ given by $\Phi(n, P) = n \cdot P = nP$ for every $n \in \mathbb{N}^*$ and every $P \in \varepsilon$ provides the corresponding versions of the preceding protocols for elliptic curves. In [Niu14] an implementation of the second protocol can be found.

(d) In [MMR07, Example 5.13], the authors illustrate the use of a semiring of order 6 to construct an example of a practical SAP. This was later cryptanalyzed in [SC11] not due to a general attack, but rather due to the structure of this ring. However, we can use the semiring of order 20 given in [MMR07, Example 5.8] to analogously define another SAP and its cryptanalysis is still an open question. This shows an example where SAP does not coincide with a traditional DLP on a semigroup and it is applicable to both preceding protocols.

(e) In [Gni14, Protocol 80], the author defines a key exchange protocol whose security is based on the SAP derived from the following semigroup action: Let S be a semiring, T a finitely generated additive subsemigroup of S and let $\mathrm{End}_+(T)$ be its (additive) endomorphism semigroup. Then the semigroup action that defines the security of this protocol is given by $\Phi : (S, T^{\mathrm{op}}) \times \mathrm{End}_+(T) \longrightarrow \mathrm{End}_+(T)$, $((s, t), f) \mapsto (x \mapsto s \cdot f(x) \cdot t)$.

Remark 5.7. Many examples of semigroup actions suitable to defining a Diffie-Hellman type key exchange protocol can be found in [Maz03]. The corresponding SAP is shown to be computationally equivalent to a DLP for some of them.

5.2.4 A Key Agreement given by a Group Action

The existence of inverses in the semigroup G acting on the set S can provide a way to agree on a common key with reduced communication overheads. Moreover, computations can be made more equally distributed among the users. We remark that in the protocols given in the two previous sections, these requirements are higher the further away the user is from the one that initialized the protocol.

Thus we assume that G is a group. The protocol is given by the following steps.

Protocol 5.8 (GSAP-3)**.** Users agree on an element $C_0 = s$ in a finite set S, a finite abelian group G, and a G-action Φ on S. For every $i = 1, \dots, n$, the user $\mathcal{U}_i$ holds a private element $g_i \in G$.

(1) For $i = 1, \dots, n-2$, user $\mathcal{U}_i$ sends to user $\mathcal{U}_{i+1}$ the message $C_i = g_i \cdot C_{i-1}$.

(2) User $\mathcal{U}_{n-1}$ computes $C_{n-1} = g_{n-1} \cdot C_{n-2}$ and broadcasts it to the other users $\{\mathcal{U}_1, \dots, \mathcal{U}_{n-2}, \mathcal{U}_n\}$.

(3) User $\mathcal{U}_n$ computes the element $g_n \cdot C_{n-1}$.

(4) For $i = 1, \dots, n-1$, user $\mathcal{U}_i$ computes $D_i = g_i^{-1} \cdot C_{n-1}$ and sends it to user $\mathcal{U}_n$.

(5) For $i = 1, \dots, n-1$, user $\mathcal{U}_n$ computes $g_n \cdot D_i$ and sends to the other users $\{\mathcal{U}_1, \dots, \mathcal{U}_{n-2}, \mathcal{U}_{n-1}\}$ the set of values $\{g_n \cdot D_1, \dots, g_n \cdot D_{n-1}, C_{n-1}\}$.

(6) For $i = 1, \dots, n-1$, user $\mathcal{U}_i$ computes $g_i \cdot (g_n \cdot D_i)$.

It is easy to see that for $i = 1, \dots, n-1$, we have that

$$C_i = \left(\prod_{j=1}^{i} g_j \right) \cdot s, \quad D_i = \left(\prod_{j=1;\ j \neq i}^{n-1} g_j \right) \cdot s,$$

and finally

$$K = C_n = \left(\prod_{j=1}^{n} g_j \right) \cdot s.$$

This follows easily from the commutativity of G and the fact that Φ is a group action.

Remark 5.9. As in Protocol 5.5, we also point out that the element C_{n-1}, which is broadcast by $\mathcal{U}_n$ in step (5) of Protocol 5.8, is needed only for future rekeying purposes.

Remark 5.10. Using the action $x \cdot s = s^x$ for $x \in \mathbb{Z}_q^*$ and $s \in S$, for a cyclic group $S = \langle g \rangle$ of order q, we get the third protocol introduced in [STW96] and [STW00], which is referred to as IKA.2 in CLIQUES [STW00]. In this case, user $\mathcal{U}_i$ sends to

user $\mathcal{U}_n$ the message $g^{\prod_{j=1,j\neq i}^{n-1} x_j}$, which is computed with the element $x_i^{-1} \mod q$, given that the x_i are selected either to be coprime with q or, as the authors suggest, q is chosen to be a prime.

An elliptic curve version is clearly also feasible. An implementation in this sense can be found in [Niu14].

5.2.5 Rekeying Operations

Another important issue in any group key management is rekeying after the initial key agreement. There exist three different situations that require a rekeying operation. The first is simply due to key caducity and the group of users remains the same. In the other two cases, we may find a new user that wishes to join the group or a user who leaves the group. In both situations it is required that the new (resp. former) user cannot access the former (resp. new) distributed key. In the following lines we describe the procedures as well as their security.

Let us start by considering the protocol GSAP-1 described in Section 5.2.1. In this case, we could simply require that a new initial key agreement is needed. However, we may shorten the rekeying process, keeping somehow the spirit of the protocol. If rekeying is due to key caducity, then user $\mathcal{U}_n$ chooses a new private element $g_n' \in G$ and defines a new sequence $f_j^n = g_n' g_n \cdot C_{j-1}$, $j = 1, \dots, n-1$, with $C_0 = s$, as is done in step (3) of GSAP-1. The rest of the users also proceed as in step (3) and recover (using their private keys as described in GSAP-1) the new key $\left(g_n' \prod_{j=1}^{n} g_j\right) \cdot s$.

In case some user, say $\mathcal{U}_i$, leaves the group, then the corresponding value f_i^n is omitted in the new message made by $\mathcal{U}_n$.

Finally, in case a user $\mathcal{U}_{n+1}$ joins the group, then user $\mathcal{U}_n$ chooses a new element g_n' and sends the message

$$\left\{ g_n' g_1 \cdot s,\ g_n' g_2 g_1 \cdot s,\ \dots,\ \left(g_n' \prod_{j=1}^{n} g_j \right) \cdot s \right\}$$

to user $\mathcal{U}_{n+1}$. Then this user starts step (3) of GSAP-1.

In the case of the protocols GSAP-2 and GSAP-3, described in Sections 5.2.2 and 5.2.4 respectively, we may use the information that every user holds after the initial key agreement to rekey very efficiently, as is suggested in [STW00]. In this case, given that every user remembers the same information, say

$$\left\{ \left(\prod_{r=2}^{n} g_r \right) \cdot s, \left(\prod_{r=1;r\neq 2}^{n} g_r \right) \cdot s, \dots, \left(\prod_{r=1;r\neq c}^{n} g_r \right) \cdot s, \dots, \left(\prod_{r=1}^{n-1} g_r \right) \cdot s \right\},$$

the rekeying process may be carried out by any one of them. Let us call this user $\mathcal{U}_c$. If rekeying is due to key caducity, then he chooses a new $g_c' \in G$, changes his private key to $g_c' g_c$ and sends the following rekeying message:

$$\left\{ \left(g_c' \prod_{r=2}^{n} g_r \right) \cdot s, \left(g_c' \prod_{r=1;r\neq 2}^{n} g_r \right) \cdot s, \dots, \left(\prod_{r=1;r\neq c}^{n} g_r \right) \cdot s, \dots, \left(g_c' \prod_{r=1}^{n-1} g_r \right) \cdot s \right\},$$

Then, every user, using his private information, recovers the new common key given by $\left(g_c' \prod_{r=1}^{n} g_r\right) \cdot s$.

In case some user leaves the group, the corresponding position in the rekeying message is omitted. If a new user joins the group, then $\mathcal{U}_c$ adds the element $\left(g_c' \prod_{r=1}^{n} g_r\right) \cdot s$ and sends the following to the new user $\mathcal{U}_{n+1}$:

$$\left\{ \left(g_c' \prod_{r=2}^{n} g_r \right) \cdot s, \dots, \left(\prod_{r=1; r\neq c}^{n} g_r \right) \cdot s, \dots, \left(g_c' \prod_{r=1}^{n-1} g_r \right) \cdot s, \left(g_c' \prod_{r=1}^{n} g_r \right) \cdot s \right\},$$

This user proceeds (in both GSAP-2 and GSAP-3) to step (5) of GSAP-3.

5.3 Security of the Key Agreements and Rekeying Operations

In [MMR07], it was pointed out that if an adversary is able to solve the SAP, then she will be able to break the two party Diffie-Hellman key exchange, i.e. solve the DHSAP. It is easy to observe that being able to solve the DHSAP allows getting the shared key in all the protocols proposed above.

Proposition 5.11. *If an adversary is able to solve the DHSAP, then she can get the shared key in GSAP-1, GSAP-2 and GSAP-3.*

Proof. This follows from the fact that the adversary can access the pair of values

- $(C_1, f_1^2) = \left(g_1 \cdot s, \left(\prod_{i=2}^{n} g_i\right) \cdot s\right)$ in GSAP-1;
- $(C_1^1, f_1^n) = \left(g_1 \cdot s, \left(\prod_{i=2}^{n} g_i\right) \cdot s\right)$ in GSAP-2;
- $(C_1, g_n \cdot D_1) = \left(g_1 \cdot s, \left(\prod_{i=2}^{n} g_j\right) \cdot s\right)$ in GSAP-3.

□

The preceding result shows, as could be expected, that the multiparty key exchange protocols do not enhance the security that the corresponding two-party protocol offers. However, as in [STW96] and [STW00], it is possible to show that increasing the number of messages does not produce any information leakage whenever the corresponding key exchange based on the SAP for two communicating parties is secure. Here we are referring to security against passive attacks; a totally different picture would arise if we assume that the attacker can control communications from and to one or more particular users, see e.g. [Sch+16].

Let $X = \{g_1, \dots, g_n\}$ be a set of elements of the semigroup G, s an element of a set S and Φ a G-action on S. Let us define the (ordered) set of elements of S

$$V_\Phi^G(s, n, X) = \left\{ \left(\prod_{j=i_1}^{i_m} g_j \right) \cdot s \;\middle|\; \{i_1, \dots, i_m\} \subsetneq \{1, \dots, n\} \right\}$$

and the value $K_\Phi^G(s, n, X) = \left(\prod_{j=1}^{n} g_j\right) \cdot s \in S$.

We point out that the messages that any adversary observes in any of the protocols is a subset of $V_\Phi^G(s, n, X)$, and the key that the users agree on is precisely $K_\Phi^G(s, n, X)$. Let us assume now that Φ is a transitive action, i.e. for every pair of elements $s, s' \in S$ there always exists a $g \in G$ such that $g \cdot s = s'$. Thus if $s \in S$ is a public element, given any two elements in S, s_1, s_2, there always exist $g_1, g_2 \in G$ such that $g_i \cdot s = s_i$, $i = 1, 2$. Let $s_3 = g_1 \cdot (g_2 \cdot s) = g_1 g_2 \cdot s$. If, given s, s_1 and s_2, it is not feasible to distinguish s_3 from a random value in polynomial time, then an induction argument like that given in [STW00, Theorem 1] allows us to show the following result.

Theorem 5.12. *Let Φ be a transitive G-action on S. Then the group key that users derive as a result of any of the protocols GSAP-1, GSAP-2 and GSAP-3 is indistinguishable in polynomial time from a random value, given only the values exchanged between users during the protocol, whenever the corresponding Diffie-Hellman protocol induced by Φ for two users satisfies this property.*

The security of the rekeying operations described in Section 5.2.5 also follows from Theorem 5.12.

5.4 Secure Group Communication based on Linear Actions

As can be observed in the protocols given in the previous section, user $\mathcal{U}_n$ plays a central role, and in two of them, every user is required to do a different number of computations and store a different number of values, depending on his proximity to $\mathcal{U}_n$. The aim of this section is twofold: On one hand, we give a similar approach to that of GSAP-3 in order to get a protocol with the same advantages that is applicable in situations where the semigroup G acting on S does not contain inverses. On the other hand, we give a new approach based on linear actions that in some cases not only significantly decreases communication overheads, but also reduces the number of rounds to just two, which will significantly enhance the efficiency.

We say that, given semigroups $(G, \cdot)$ and $(S, +)$, an action $\Phi : G \times S \longrightarrow S$ is linear in case $g \cdot (s + s') = g \cdot s + g \cdot s'$.

The following protocol is a modification of GSAP-3 for a linear G-action Φ on S, but instead of requiring G to be a group, we require this of S. We get a similar protocol that is also an extension of Diffie-Hellman to the multiparty case.

Protocol 5.13 (GSAP-3′)**.** Users agree on an element s in a finite group S, a finite abelian semigroup G, and a linear G-action Φ on S. For every $i = 1, \ldots, n$, the user $\mathcal{U}_i$ holds a private element $g_i \in G$.

(1) For $i = 1, \ldots, n-2$, user $\mathcal{U}_i$ sends to user $\mathcal{U}_{i+1}$ the message $C_i = g_i \cdot C_{i-1}$.

(2) User $\mathcal{U}_{n-1}$ computes $C_{n-1} = g_{n-1} \cdot C_{n-2}$ and broadcasts it to the other users $\{\mathcal{U}_1, \dots, \mathcal{U}_{n-2}, \mathcal{U}_n\}$.

(3) User $\mathcal{U}_n$ computes the element $g_n \cdot C_{n-1}$.

(4) For $i = 1, \dots, n-1$, user $\mathcal{U}_i$ computes $D_i = C_{n-1} - g_i \cdot s$ and sends it to user $\mathcal{U}_n$.

(5) For $i = 1, \dots, n-1$, user $\mathcal{U}_n$ computes $g_n \cdot D_i$ and sends to the other users $\{\mathcal{U}_1, \dots, \mathcal{U}_{n-2}, \mathcal{U}_{n-1}\}$ the set of values $\{g_n \cdot D_1, \dots, g_n \cdot D_{n-1}, g_n \cdot D_n\}$ and his public key $g_n \cdot s$, where $D_n = C_{n-1} - g_n \cdot s$.

(6) For $i = 1, \dots, n-1$, user $\mathcal{U}_i$ computes $g_i \cdot (g_n \cdot s) + g_n \cdot D_i$.

Theorem 5.14. *After protocol GSAP-3′, the users $\mathcal{U}_1, \dots, \mathcal{U}_n$ share a common key given by $\left(\prod_{i=1}^{n} g_i\right) \cdot s$.*

Proof. This follows from the linearity of the action Φ:

$$g_i \cdot (g_n \cdot s) + g_n \cdot D_i = g_i g_n \cdot s + g_n \cdot \left(\left(\prod_{r=1}^{n-1} g_r \right) \cdot s - g_i \cdot s \right)$$
$$= \left(\prod_{r=1}^{n} g_r \right) \cdot s,$$

since $g_i \cdot e = e$, e being the neutral element in S, and $-(g_i \cdot s) = g_i \cdot (-s)$, again by the linearity of the action. □

Example 5.15. (a) Given a cyclic group $(S, \cdot)$ of order q generated by g, the action $\Phi : \mathbb{N}^* \times S \longrightarrow S$ defined by $\Phi(x, s) = s^x$ is clearly linear, so the above argument applies. D_i assumes the form $g^{\prod_{j=1}^{n-1} x_j} g^{-x_i}$.

(b) If ε is the group of points of an elliptic curve, then ε is a $\mathbb{Z}$-module via the linear action $\Phi(k, P) = kP$ for every $k \in \mathbb{Z}$ and $P \in \varepsilon$. D_i assumes the form $\left(\prod_{i=1}^{n-1} k_j\right) P - k_i P$.

(c) Let us introduce an example where the preceding protocols can be run over a module structure. Let us recall from [CNT14] the following ring:

$$E_p^{(m)} = \left\{ [a_{ij}] \in \mathrm{Mat}_{m \times m}(\mathbb{Z}) \;\middle|\; a_{ij} \in \mathbb{Z}_{p^i} \text{ if } i \leq j, \text{ and } a_{ij} \in p^{i-j}\mathbb{Z}_{p^i} \text{ if } i > j \right\},$$

with addition and multiplication defined, respectively, as follows

$$[a_{ij}] + [b_{ij}] = [(a_{ij} + b_{ij}) \bmod p^i]$$
$$[a_{ij}] \cdot [b_{ij}] = \left[\left(\sum_{k=1}^{m} a_{ik} b_{kj} \right) \bmod p^i \right].$$

Here, $\mathrm{Mat}_{m\times m}(\mathbb{Z})$ denotes the set of $m \times m$ matrices with entries in $\mathbb{Z}$, and $p^r\mathbb{Z}_{p^s}$ denotes the set $\{p^r u \mid u \in \{0,\dots,p^s-1\}\} \subset \mathbb{Z}$ for positive integers r and s. This ring is clearly non-commutative and its product defines an action of the multiplicative semigroup $E_p^{(m)}$ on the set $\mathbb{Z}_p \times \mathbb{Z}_{p^2} \times \cdots \times \mathbb{Z}_{p^m}$. However, to ensure that the key exchange works, we need that the elements in the semigroup commute. In this non-commutative setting, this may be achieved by considering that the selected elements in the semigroup $E_p^{(m)}$ are of the form $\sum_{i=0}^{r} C_i M^i$, such that for every $i = 0,\dots,r$, C_i is in the center Z of $E_p^{(m)}$ and $M \in E_p^{(m)}$ is a public element such that its set of powers is large enough. In other words, if we denote the set of elements of this form by $Z[M]$, then we are using for G the multiplicative subsemigroup $Z[M]$ of $E_p^{(m)}$.

From [CL16, Theorem 2] we can deduce conditions on the public information that will be sent in order to prevent an attacker from solving the SAP in the subsemigroup of $Z[M]$ given by the center Z of the ring, with cardinality p^m (cf. [CNT14]). Thus if M has high order, i.e. M is such that the least integer n satisfying $M^{k+n} = M^k$ for every sufficiently large k is high, we will obtain that $Z[M]$ is big enough.

Note that our aim in this paper is not to prove the hardness of the SAP for this particular example, but rather to present protocols which rely on the hardness of the SAP in a particular scenario once it has been established there. The non-commutative scenario in particular may present hidden vulnerabilities, as was shown in recent cryptanalyses, e.g. [KY12; Mic15], although these seem not to directly apply in this setting. For example [KY12] introduces a cryptanalysis for the case of two users when the ring $E_p^{(m)}$ acts on itself, which can be countered by choosing p and m appropriately in order to avoid the existence of inverses [CNT14]. In the case of [Mic15], the cryptanalysis requires building a system of equations, which does not seem to be straightforward in this new setting of $Z[M]$. In [Maz03, Proposition 3.9], it is asserted that if the commutative semigroup has a big number of invertible elements, then it is possible to develop a square root attack to the SAP. Again we point out that $E_p^{(m)}$ could be chosen in order to avoid this attack.

Given that both $\left(\prod_{i=1}^{n-1} g_i\right) \cdot s$ and $g_n \cdot s$ are public we immediately get the following.

Proposition 5.16. *If an adversary is able to solve the DHSAP, then she can get the shared key in GSAP-3′.*

Let us recall from [MMR07] that given any G-action Φ on S, we can easily define an ElGamal type of public key cryptosystem. We define the following ElGamal type of protocol.

(1) Alice and Bob publicly agree on an element $s \in S$.

(2) Bob chooses $b \in G$ and computes $b \cdot s$. Bob's private key is b, his public key is $b \cdot s$.

(3) If Alice wants to send the message $m \in S$ to Bob, then she gets Bob's public key $b \cdot s$.

(4) Alice chooses randomly $a \in G$ and computes $a \cdot s$ and $a \cdot (b \cdot s)$.

(5) Alice sends to Bob the pair $(c, d) = (a \cdot s,\ m + a \cdot (b \cdot s))$.

(6) Bob recovers $m = d - b \cdot c = m + a \cdot (b \cdot s) - b \cdot (a \cdot s)$, given that S has a group structure.

It can be easily observed that solving the DHSAP is equivalent to breaking the preceding algorithm: If given the public information

$$(s,\ a \cdot s,\ b \cdot s,\ m + ab \cdot s)$$

one is able to get m, then the input $(s, a \cdot s, b \cdot s, e)$, for $e \in S$ the neutral element, produces $-(ab \cdot s)$, which solves the DHSAP. Conversely, given Bob's public key $b \cdot s$ and the pair $(a \cdot s, m + a \cdot (b \cdot s))$, one can use $ab \cdot s$ from the DHSAP to recover m.

Now using the above we are able to show the security of GSAP-3′.

Theorem 5.17. *The group key that users derive as a result of GSAP-3′ is indistinguishable in polynomial time from a random value whenever the corresponding Diffie-Hellman protocol induced by Φ for two users also satisfies this property.*

Proof. Given that both $C_{n-1} = \left(\prod_{i=1}^{n-1} g_i\right) \cdot s$ and $D_i = C_{n-1} - g_i \cdot s$ are public, an adversary is able to get all the public values $g_i \cdot s$, $i = 1, \dots, n$. Now user $\mathcal{U}_n$ sends the message $\{g_n \cdot D_i\}_{i=1}^{n-1}$ jointly with $g_n \cdot s$, in other words, due to linearity of Φ, user $\mathcal{U}_n$ sends a "a family of pairs", $i = 1, \dots, n$,

$$\left(g_n \cdot s,\ -g_n \cdot (g_i \cdot s) + g_n \cdot \left(\prod_{j=1}^{n-1} g_j\right) \cdot s\right),$$

which can be seen as a set of ElGamal encryptions of the message

$$\left(\prod_{i=1}^{n} g_i\right) \cdot s = g_n \cdot \left(\prod_{i=1}^{n-1} g_i\right) \cdot s$$

using the public keys $g_i \cdot s$ for $i = 1, \dots, n$. Alternatively, one can consider the pairs

$$\left(g_i \cdot s,\ -g_n \cdot (g_i \cdot s) + g_n \cdot \left(\prod_{j=1}^{n-1} g_j\right) \cdot s\right),$$

which can also be seen, given the commutativity in G, as a set of ElGamal encryptions of the message

$$\left(\prod_{i=1}^{n} g_i\right) \cdot s = g_n \cdot \left(\prod_{i=1}^{n-1} g_i\right) \cdot s$$

using the public key $g_n \cdot (-s)$, and the g_i as random numbers, for $i = 1, \ldots, n$.

Thus, as we pointed out above, given the equivalence of the security of the ElGamal type of public key cryptosystem and the DHSAP, the result follows. □

The rekeying process in this setting is analogous to that described in Section 5.2.5 for protocols GSAP-2 and GSAP-3.

We first note that every user remembers the following keying information:

$$\{g_n \cdot D_1,\ \ldots,\ g_n \cdot D_{n-1},\ g_n \cdot D_n\}.$$

In case of key caducity, user $\mathcal{U}_c$ for some $c = 1, \ldots, n$ chooses a new element $g'_c \in G$, computes a new key given by $\left(g'_c \prod_{i=1}^{n} g_i\right) \cdot s$ and his keying information $\left(g'_c \prod_{i=1}^{n} g_i\right) \cdot s - (g'_c)^2 g_c g_n \cdot s$ and broadcasts the following message:

$$\left\{ g'_c \cdot (g_n \cdot D_1),\ \ldots,\ \left(g'_c \prod_{i=1}^{n} g_i\right) \cdot s - (g'_c)^2 g_c g_n \cdot s,\ \ldots, \right.$$
$$\left. g'_c \cdot (g_n \cdot D_{n-1}),\ g'_c \cdot (g_n \cdot D_n) \right\},$$

jointly with the value $g'_c \cdot (g_n \cdot s)$. User $\mathcal{U}_c$ changes his private information to $g_c g'_c$.

In case rekeying is due to some user leaving the group, then the corresponding value is omitted in the above message.

Finally, let us assume that $\mathcal{U}_{n+1}$ joins the group. The process corresponds in this case to something similar to a "double rekeying" as above. First, $\mathcal{U}_c$ sends to $\mathcal{U}_{n+1}$

$$\left\{ g'_c \cdot (g_n \cdot D_1), \ldots, g'_c \cdot \left(\prod_{i=1}^{n} g_i\right) \cdot s - (g'_c)^2 g_c g_n \cdot s, \ldots, \right.$$
$$\left. g'_c \cdot (g_n \cdot D_{n-1}), g'_c \cdot (g_n \cdot D_n), g'_c \cdot \left(\prod_{i=1}^{n} g_i\right) \cdot s \right\}$$

jointly with the value $g'_c \cdot (g_n \cdot s)$. Then, $\mathcal{U}_{n+1}$ broadcasts a rekeying message given by

$$\left\{ g_{n+1} g'_c \cdot (g_n \cdot D_1), \ldots, g'_c \cdot \left(\prod_{i=1}^{n+1} g_i\right) \cdot s - g_{n+1} (g'_c)^2 g_c g_n \cdot s, \ldots, \right.$$
$$g_{n+1} g'_c \cdot (g_n \cdot D_{n-1}), g_{n+1} g'_c \cdot (g_n \cdot D_n),$$
$$\left. g'_c \cdot \left(\prod_{i=1}^{n+1} g_i\right) \cdot s - g_{n+1}^2 g'_c g_n \cdot s \right\}$$

jointly with the value $g_{n+1}g_c'g_n \cdot s$.

Security of these processes can be shown with a similar argument as in Theorem 5.17.

A more symmetrical use of linear actions is the following protocol, which decreases the number of rounds to just two, but which is only applicable in some cases.

Protocol 5.18 (GSAP-4)**.** Users agree on an element s in a finite abelian semigroup S, a finite abelian semigroup G, and a linear G-action Φ on S. For every $i = 1, \dots, n$, the user $\mathcal{U}_i$ holds a private element $g_i \in G$.

(1) For every $i = 1, \dots, n$, user $\mathcal{U}_i$ makes public $g_i \cdot s$.

(2) For some $j = 1, \dots, n$, user $\mathcal{U}_j$ computes and makes public

$$D_i = g_j \cdot \sum_{r \neq j,i} (g_r \cdot s), \ i \neq j, \ i = 1, \dots, n.$$

(3) For every $i = 1, \dots, n$, $i \neq j$, user $\mathcal{U}_i$ computes $D_i + g_i \cdot (g_j \cdot s)$. User $\mathcal{U}_j$ computes $g_j \cdot \sum_{r \neq j} (g_r \cdot s)$.

Theorem 5.19. *After protocol GSAP-4, the users $\mathcal{U}_1, \dots, \mathcal{U}_n$ share a common key given by $g_j \cdot \sum_{r \neq j} (g_r \cdot s)$.*

Proof. For every $i = 1, \dots, n$, $i \neq j$,

$$\begin{aligned} D_i + g_i \cdot (g_j \cdot s) &= g_j \cdot \sum_{r \neq j,i} (g_r \cdot s) + g_i \cdot (g_j \cdot s) \\ &= g_j \cdot \sum_{r \neq j,i} (g_r \cdot s) + g_i g_j \cdot s \\ &= g_j \cdot \sum_{r \neq j,i} (g_r \cdot s) + g_j g_i \cdot s \\ &= g_j \cdot \sum_{r \neq j,i} (g_r \cdot s) + g_j \cdot (g_i \cdot s) \\ &= g_j \cdot \sum_{r \neq j} (g_r \cdot s). \end{aligned}$$

□

Example 5.20. (a) Let us consider again a cyclic group S of order q generated by g, with the action $\Phi : \mathbb{N}^* \times S \longrightarrow S$ given by $\Phi(x, s) = s^x$. Then GSAP-4 implies sharing a key of the form $K = g^{k_j \sum_{r=1, r \neq j}^{n} k_r}$. An adversary can access the messages

$$D_i = g^{k_j \sum_{r=1; r \neq i,j}^{n} k_r}, \text{ for } i = 1, \dots, n, \ i \neq j,$$

from which she can compute $\prod_{r=1,r\neq j}^{n} D_r = K^{n-2}$. In the case where the order q of S is known, the adversary can now recover the key K from K^{n-2} by inverting $n-2$ modulo q. This is in particular the case where S is a subgroup of a finite field, or where it is the group of points of an elliptic curve. However, we can avoid this weakness by adding some authentication information as is done in [AST00].

(b) Let $m = pq$ with p and q two large primes and let $G = \mathbb{Z}^*_{(p-1)(q-1)}$. Then the action $\Phi : G \times \mathbb{Z}_m \longrightarrow \mathbb{Z}_m$ given by $\Phi(x, g) = g^x \mod m$ shows an example where the above attack cannot be developed unless the adversary is able to factor m. The shared key in this case is of the form $g^{k_j \sum_{i=1,i\neq j}^{n} k_i} \mod m$.

(c) We recall that a semiring R is a semigroup with respect to both addition and multiplication and the distributive laws hold. It is also understood that a semiring is commutative with respect to addition and the existence of neutral elements is not required, although some authors do require it. Then, given a semiring R, a left R-semimodule M is an abelian semigroup with an action $\Phi : R \times M \longrightarrow M$, $\Phi(r, m) = rm$, satisfying $r(sm) = (rs)m$, $(r + s)m = rm + sm$ and $r(m + n) = rm + rn$ for all $r, s \in R$ and $m, n \in M$. Thus, based on the previous two examples, we can assert in general that any semimodule S over a semiring R fits with GSAP-4 and the shared key is of the form $k_j\left(\sum_{r=1,r\neq j}^{n} k_r\right)s$ for $k_i \in R$, $i = 1, \ldots, n$ private and $s \in S$ public.

Remark 5.21. Due to the attack shown in example (a), the hardness of the Diffie-Hellman problem is not enough to show security in this case. We leave it as an open question whether the hardness of factoring would be enough to do so for example (b).

Remark 5.22. We can also give protocols based on two-sided actions. To this end we recall that given a semiring S, right S-semimodules are defined dually to left ones. Then, given two semirings R and S, an (R, S)-bisemimodule M is both a left R-semimodule and a right S-semimodule such that $(rm)s = r(ms)$ for every $r \in R$, $m \in M$ and $s \in S$.

Now we are able to provide key exchange protocols similar to those given in the previous sections based on two-sided linear actions over a (R, S)-bisemimodule M. In the case of GSAP-3′, since we need the existence of inverses with respect to addition in M, we may suppose that M has an (R, S)-bimodule structure for some rings R and S.

5.5 Further Key Agreements based on Linear Actions

We now present two variants of GSAP-4 that avoid the attack described in Example 5.20 (a). They were published in [Lóp+16]. These protocols have the following two desirable properties: On one hand, the key is obtained in a distributed key agreement with just two rounds. On the other hand, the rekeying protocol

is developed by means of a single message. The protocols extend naturally the Diffie-Hellman key exchange as well and we show that their security is based on the difficulty of the DDH problem.

Let us start by establishing the general setting for the following protocols. Participants in the communication process will again be given by the set $\{\mathcal{U}_1, \dots \mathcal{U}_n\}$. The users agree on an abelian semigroup $(G, \cdot)$, an abelian group $(S, +)$, and a linear semigroup action $\Phi : G \times S \longrightarrow S$. They also agree on a base element $s \in S$.

Every participant $\mathcal{U}_i$ holds two pairs of private-public keys, say $(g_i, g_i \cdot s)$ and $(x_i, x_i \cdot s)$. One of these users is chosen to be the group controller, whom we will denote by $\mathcal{U}_{c_1}$, for some c_1 in the set $\{1, \dots, n\}$. He will be in charge of sending the initial keying information as well as the following rekeying messages in case we wish to define a centralized protocol. However, as we will see in the following section, the character of the protocol can change from centralized to distributed (and vice versa) at any point of the following rekeying stages. The protocol that describes the initial key agreement is given by the following steps.

Protocol 5.23.

First Round:

(1) Every user $\mathcal{U}_i$, $i \neq c_1$, publishes the pair $(g_i \cdot s, x_i \cdot s)$,

(2) The group controller $\mathcal{U}_{c_1}$ computes the key $K_1 = g_{c_1} \cdot \sum_{j=1, j\neq c_1}^{n} (g_j \cdot s)$.

(3) The group controller takes two new elements g'_{c_1}, $x'_{c_1} \in G$ that become his new private information.

Second Round:

(4) Every user $\mathcal{U}_i$, $i \neq c_1$, computes $\sum_{j=1, j\neq c_1, i}^{n} (g_j \cdot s)$ and sends this value to $\mathcal{U}_{c_1}$.

(5) The group controller $\mathcal{U}_{c_1}$ broadcasts the keying message

$$\{Y_{1,1}, \dots, Y_{1,c_1}, \dots, Y_{1,n}, R_1, S_1\},$$

where

$$Y_{1,i} = \left(g_{c_1} \cdot \sum_{j=1; j\neq c_1, i}^{n} (g_j \cdot s) \right) - (x_{c_1} \cdot (x_i \cdot s))$$

for $i = 1, \dots, n$, $i \neq c_1$, $Y_{1,c_1} = K_1 - g'_{c_1} \cdot (g_{c_1} \cdot s) - x'_{c_1} \cdot (x_{c_1} \cdot s)$, and $R_1 = g_{c_1} \cdot s$ and $S_1 = x_{c_1} \cdot s$.

(6) Every user $\mathcal{U}_i$, $i \neq c_1$, computes $K_{1,i} = Y_{1,i} + x_i \cdot S_1 + g_i \cdot R_1$.

The proof of the following lemma is straightforward and shows the correctness of the protocol.

Lemma 5.24. *$K_{1,i} = K_1$ for every $i = 1, \dots, n$, $i \neq c_1$.*

Remark 5.25. Let us assume that the number of users is $n = 2$ and that $x_i \cdot s = 0$ for $i = 1, 2$, where $0 \in S$ is the neutral element. Now if $\mathcal{U}_1$ makes public $g_1 \cdot s$ in the first round, $\mathcal{U}_2$ will send the keying message $\{0, R_1 = g_2 \cdot s\}$ in the second round. Thus our protocol is a natural extension of the semigroup Diffie-Hellman key exchange (Protocol 5.1).

It can be observed in the preceding protocol that user $\mathcal{U}_{c_1}$ bears most of the workload. The protocol is designed in such a way that every node publishes just a pair of public keys, while $\mathcal{U}_{c_1}$ computes what is required for the first keying. This could be the case when $\mathcal{U}_{c_1}$ is a server that processes the pieces of information transmitted by every user. However, in case every user has similar capabilities, we can slightly modify the preceding protocol and distribute the computation requirements. As previously, every user holds a pair of private keys (r_i, x_i).

Protocol 5.26.

First Round:

(1) Every user $\mathcal{U}_i$, $i \neq c_1$, publishes his public key $g_i \cdot s$,

(2) The group controller $\mathcal{U}_{c_1}$ computes the key $K_1 = g_{c_1} \cdot \sum_{j=1, j\neq c_1}^{n} (g_j \cdot s)$.

(3) The group controller takes two new elements $g'_{c_1}, x'_{c_1} \in G$ that become his new private information.

Second Round:

(4) Every user $\mathcal{U}_i$, $i \neq c_1$, computes $\sum_{j=1, j\neq c_1, i}^{n} (g_j \cdot s) - x_i \cdot s$ and sends this value to $\mathcal{U}_{c_1}$.

(5) The group controller $\mathcal{U}_{c_1}$ broadcasts the keying message

$$\{Y_{1,1}, \ldots, Y_{1,c_1}, \ldots, Y_{1,n}, R_1\},$$

where

$$Y_{1,i} = \left(g_{c_1} \cdot \sum_{j=1; j\neq c_1, i}^{n} (g_j \cdot s) \right) - (g_{c_1} \cdot (x_i \cdot s))$$

for $i = 1, \ldots, n$, $i \neq c_1$, $Y_{1,c_1} = K_1 - g'_{c_1} \cdot (g_{c_1} \cdot s) - x'_{c_1} \cdot (g_{c_1} \cdot s)$, and $R_1 = g_{c_1} \cdot s$.

(6) Every user $\mathcal{U}_i$, $i \neq c_1$, computes $K_{1,i} = Y_{1,i} + x_i \cdot R_1 + g_i \cdot R_1$.

We will now state the security of the preceding protocols. To this end let us now recall the following definition.

Definition 5.27. [BD05, Definition 2.2] Let $\mathcal{P}$ be a group key exchange protocol and $\mathcal{A}$ a passive adversary. Assume that $\mathcal{A}$ has witnessed polynomially-many instances of $\mathcal{P}$ and let K be the key output by the last instance.

We will say that $\mathcal{P}$ guarantees secrecy if $\mathcal{A}$ cannot distinguish K from a random bit string of the same length with probability better than $1/2 + \varepsilon$, where ε is negligible.

Theorem 5.28. *If the DDH problem is intractable, then Protocols 5.23 and 5.26 provide secrecy.*

Proof. We observe that we can see the broadcast message in Protocol 5.23 as a multiple ElGamal type of encryption in the following way. For $i \neq c_1$ we first encrypt K_1 using the public value $g_i \cdot s$ and g_{c_1} as random parameter, obtaining $(X_{1,i}, R_1) = \left(\left(g_{c_1} \sum_{j \neq c_1, i} g_j\right) \cdot s, g_{c_1} \cdot s\right)$ and then we encrypt X_i using the public key $x_i \cdot s$ and x_{c_1} as a random parameter, obtaining the pair $(Y_{1,i}, S_1)$.

The case of Y_{1,c_1} is analogous using the elements $g'_{c_1} \cdot s$ and $x'_{c_1} \cdot s$, that, although unknown to a passive adversary, could also be made public.

Now using Lemma 1 and Theorem 1 of [BBS02], we can deduce the claim.

The security of Protocol 5.26 follows similarly. □

5.5.1 Rekeying

The following protocol shows the rekeying operation after $t-1$ rekeying rounds have already occurred. We denote by K_t the last common key shared by the group. The user in charge of the t-th rekeying will be user $\mathcal{U}_{c_t}$, distinct from the preceding controller, and thus, rekeying information of this will be needed. Without loss of generality, we may assume that the precedent controller was user $\mathcal{U}_{c_1}$ and that the last rekeying message is given by

$$\{Y_{t-1,1}, \ldots, Y_{t-1,c_1}, \ldots, Y_{t-1,n}, R_{t-1}, S_{t-1}\},$$

where $Y_{t-1,c_1} = K_{t-1} - g'_{c_1} \cdot (g_{c_1} \cdot s) - x'_{c_1} \cdot (x_{c_1} \cdot s)$.

The protocol described here applies to the case where the set of users remain the same. If users leave the group, the operation can be done the same way, but with the entries corresponding to the leaving users removed from the rekeying message.

Protocol 5.29.

(1) User $\mathcal{U}_{c_t}$ computes two new elements g'_{c_t} and $x'_{c_t} \in G$ that become his new private information.

(2) User $\mathcal{U}_{c_t}$ computes the new session key $g'_{c_t} \cdot K_{t-1}$.

(3) User $\mathcal{U}_{c_t}$ broadcasts the rekeying message

$$\{Y_{t,1}, \ldots, Y_{t,c_t}, \ldots, Y_{t,n}, R_t, S_t\},$$

where $Y_{t,i} = g'_{c_t} \cdot Y_{t-1,i}$ for $i \neq c_t$, $Y_{t,c_t} = K_t - g'_{c_t} \cdot (g'_{c_t} \cdot R_{t-1}) - g'_{c_t} \cdot (x'_{c_t} \cdot S_{t-1})$, and $R_t = g'_{c_t} \cdot R_{t-1}$ and $S_t = g'_{c_t} \cdot S_{t-1}$.

(4) Every user $\mathcal{U}_i$, $i \neq c_t$, computes $K_{t,i} = Y_{t,i} + x_i \cdot S_t + g_i \cdot R_t$.

If instead users $\mathcal{U}_{n+j}$ for $j = 1, \ldots, l$ wish to join the group, they proceed according to the following protocol.

Protocol 5.30.

(1) Every new user $\mathcal{U}_{n+j}$, $j = 1, \dots, l$, sends a petition to user $\mathcal{U}_{c_t}$ jointly with the pair $g_{n+j} \cdot R_{t-1}$, $x_{n+j} \cdot S_{t-1}$, where $g_{n+j}, x_{n+j} \in G$ is the user $\mathcal{U}_{n+j}$'s private information.

(2) User $\mathcal{U}_{c_t}$ computes two new elements $g'_{c_t}, x'_{c_t} \in G$ that become his new private information.

(3) User $\mathcal{U}_{c_t}$ computes the new key $K_t = g'_{c_t} \cdot \left(K_{t-1} + \sum_{i=1}^{l}(g_{n+i} \cdot R_{t-1})\right)$.

(4) User $\mathcal{U}_{c_t}$ broadcasts the rekeying message

$$\{Y_{t,1}, \dots, Y_{t,c_t}, \dots, Y_{t,n}, Y_{t,n+1}, \dots, Y_{t,n+l}, R_t, S_t\}$$

where $Y_{t,i} = g'_{c_t} \cdot \left(Y_{t-1,i} + \sum_{j=1}^{l}(r_{n+j} \cdot R_{t-1})\right)$, for $i = 1, \dots, n$, $i \neq c_t$,
$Y_{t,c_t} = K_{t-1} - g'_{c_t} g'_{c_t} \cdot R_{t-1} - g'_{c_t} x'_{c_t} \cdot S_{t-1}$,
$Y_{t,i} = K_{t-1} - g'_{c_t} g_i \cdot R_{t-1} - g'_{c_t} x_i \cdot S_{t-1}$, for $i = n+1, \dots, n+l$,
$R_t = g'_{c_t} \cdot R_{t-1}$ and $S_t = g'_{c_t} \cdot S_{t-1}$.

(5) Every user $\mathcal{U}_i$ computes $K_{t,i} = Y_{t,i} + x_i \cdot S_t + g_i \cdot R_t$, $i = 1, \dots, n+l$, $i \neq c_t$.

5.6 An Active Attack on GSAP-3

In this section, we will show an active attack on the protocol GSAP-3 of [STW96] and our variant Protocol 5.8, which requires control of the communications of two particular parties for only the duration of the key exchange. That is, unlike in a regular man-in-the-middle attack, it is not necessary for the attacker to control the communications after the key exchange in order to translate messages, since all users are made to agree on the same key. This attack was published in [Sch+16].

Although it is not possible for the attacker to keep a copy of the key after the users initiate rekeying operations, we will show how she can avoid being noticed at that point.

5.6.1 The Attack

Recall that after the execution of GSAP-3 (Protocol 5.8), we have for $i = 1, \dots, n-1$

$$C_i = \left(\prod_{j=1}^{i} g_j\right) \cdot s, \quad D_i = \left(\prod_{j=1;\ j\neq i}^{n-1} g_j\right) \cdot s,$$

and finally

$$K = C_n = \left(\prod_{j=1}^{n} g_j\right) \cdot s.$$

We describe an active attack on GSAP-3. Suppose that the attacker $\mathcal{M}$ wants the users $\mathcal{U}_1, \ldots, \mathcal{U}_n$ to agree on a shared key as usual, except that she is in possession of the key as well.

In order to carry out our attack, $\mathcal{M}$ needs to have full control over the communication of the users $\mathcal{U}_{n-1}$ and $\mathcal{U}_n$ for the duration of the key exchange.

In the beginning, $\mathcal{M}$ chooses her own secret group element $\hat{g} \in G$. She then proceeds as follows:

(a) Step (1) of GSAP-3 is carried out as usual.

(b) $\mathcal{M}$ intercepts the broadcast of $\mathcal{U}_{n-1}$ during step (2) and remembers the value C_{n-1}. At this point, all users except for $\mathcal{U}_{n-1}$ are sitting in step (2), waiting for the broadcast that was halted.

(c) $\mathcal{U}_{n-1}$ proceeds to step (4), where he sends $g_{n-1}^{-1} \cdot C_{n-1} = C_{n-2}$ to $\mathcal{U}_n$. This is also intercepted by $\mathcal{M}$. $\mathcal{U}_{n-1}$ is now waiting in step (5).

(d) $\mathcal{M}$ now makes $\mathcal{U}_n$ believe that he received the broadcast of step (2), but actually sends him $\hat{g} \cdot C_{n-1}$. At this point, $\mathcal{U}_n$ computes the shared key $K = g_n \hat{g} \cdot C_{n-1}$ and waits in step (4).

(e) $\mathcal{M}$ now sends to $\mathcal{U}_n$ the values $\{m_1, \ldots, m_{n-3}, C_{n-2}, C_{n-1}\}$, pretending that they were sent by the other users in step (4). The m_i are random elements of the orbit $G \cdot s$.

(f) In step (5), $\mathcal{U}_n$ sends back, among others, the values $g_n \cdot C_{n-2}$ and $g_n \cdot C_{n-1}$, which $\mathcal{M}$ intercepts. The user $\mathcal{U}_n$ is now finished, and $\mathcal{M}$ can compute the shared key $K = \hat{g} g_n \cdot C_{n-1}$.

(g) Until now, $\mathcal{U}_1, \ldots, \mathcal{U}_{n-2}$ have been waiting for the broadcast in step (2), which $\mathcal{M}$ now provides in the form of $g_n \cdot C_{n-1}$.

(h) $\mathcal{U}_i$, $i = 1 \ldots, n-2$, go to step (4) and send back $g_i^{-1} g_n \cdot C_{n-1}$, which $\mathcal{M}$ intercepts.

(i) In step (5), $\mathcal{M}$ broadcasts to $\mathcal{U}_i$, $i = 1, \ldots, n-2$, the message

$$\{\hat{g} g_1^{-1} g_n \cdot C_{n-1},\ \hat{g} g_2^{-1} g_n \cdot C_{n-1},\ \ldots,\ \hat{g} g_{n-1}^{-1} g_n \cdot C_{n-1},\ g_n \cdot C_{n-1}\}$$

User $\mathcal{U}_{n-1}$ is sent the same message, but the last element, $g_n \cdot C_{n-1}$ is substituted by C_{n-1}.

(j) The users $\mathcal{U}_1, \ldots, \mathcal{U}_{n-2}$ now all compute the shared secret $K = g_i \hat{g} g_i^{-1} g_n \cdot C_{n-1}$.

Let us make some comments on the attack introduced above. First, we can observe that at the end of this procedure, all users as well as the attacker share the same key

$$K = \left(\prod_{j=1}^{n} g_j \right) \cdot (\hat{g} \cdot s).$$

Any passive observer will still be unable to determine the key, for the same reason that the original protocol is secure against passive attacks, whenever the action is transitive and the Diffie-Hellman problem is hard (see Section 5.3 and [STW00, Theorem 2.1]).

The attacker's secret $\hat{g}$ is not strictly required for the attack to work, but without it, the users may notice that something is amiss. Namely, in step (e), if we leave out $\hat{g}$, the user $\mathcal{U}_n$ may notice that $\mathcal{M}$ sent the same value C_{n-1} as in step (d). Similarly, in step (i), the other users could notice that the attacker just returned their transmission from (h). Using $\hat{g}$, however, the users should be unable to tell the difference between a regular execution of the protocol and the attack, again as a consequence of [STW00, Theorem 2.1].

As in the original protocol, the broadcast element $g_n \cdot C_{n-1}$ is added at the end of the message in step (i) in view of future rekeying operations and is not needed by any of the users $\mathcal{U}_1, \dots, \mathcal{U}_{n-2}$ to recover the shared key. Note that users $\mathcal{U}_i$, $i = 1, \dots, n-2$, expect that the last element of the message sent in step (i) is the one broadcast in step (2) of the protocol, which the attacker substitutes precisely by $g_n \cdot C_{n-1}$. In the case of user $\mathcal{U}_{n-1}$, who is also expecting the element sent in step (2) of the protocol, the element that $\mathcal{M}$ sends in step (b) is C_{n-1}. If this is not satisfied, the users might notice that something is wrong.

Remark 5.31. We want to point out that since the protocol does not include any authentication, it is always possible for an active attacker to do a simple man-in-the-middle attack. However, in that case, the attacked user ends up with a different key from the rest of the group. The attacker must therefore maintain control over that user's communications and translate messages between the attacked user and the rest of the group. In our attack however, the attacker needs to maintain control only for the duration of the key exchange, since all users end up with the same key. Afterwards, the attacker can passively listen to the conversation.

5.6.2 An Exit Strategy

After the attack of Section 5.6, the attacker $\mathcal{M}$ shares the key with the users $\mathcal{U}_1, \dots, \mathcal{U}_n$ and can listen in on their conversation without any further active measures. However, at some point after that, the users may wish to execute a rekeying operation, which is to say a key refreshment, the addition of a new member to the group, etc. as described in [STW00, Section 5]. After this point, the attacker can certainly no longer listen to the conversation. Even worse, the values the users remember from step (5) of the protocol are substantially different from normal, and any key refresh operation will thus fail completely, alerting the users about the attack.

In what follows, we will describe how the attacker can avoid being noticed by forging key refresh operations herself, assuming that any user may initiate a key refreshment at any time.

First, we recall the key refresh operation after a regular execution of GSAP-3, see Section 5.2.5 and [STW00, Section 5.6]. Suppose user $\mathcal{U}_c$ wishes to initiate a key refreshment. He remembers from step (5) of the key agreement protocol the values $\{E_1, \dots, E_n\}$, where $E_k = \left(\prod_{j=1, j\neq k}^{n} g_j\right) \cdot s$, $k = 1, \dots, n$. He picks a new secret $g_c' \in G$ and broadcasts

$$\{g_c' \cdot E_1, \dots, g_c' \cdot E_{c-1}, E_c, g_c' \cdot E_{c+1}, \dots, g_c' \cdot E_n\}.$$

Now, all users can compute the new key $g_c' \cdot C_n = g_c' \cdot \left(\prod_{j=1}^{n} g_j\right) \cdot s$. User $\mathcal{U}_c$ also replaces his own secret with $g_c' g_c$, and everyone replaces the information remembered from step (5) with this new broadcast.

Remark 5.32. One important detail to note is that when $\mathcal{U}_c$ initiates the key refreshment, the value E_c he sends in position c is unchanged and already known to the other users. Hence, if $\mathcal{M}$ wishes to forge a key refreshment coming from $\mathcal{U}_c$, she has to make sure that each user receives in position c the value he previously held there. Otherwise, the attack could be discovered.

Suppose now that the attacker $\mathcal{M}$ has just executed the attack from Section 5.6. Instead of $\{E_1, \dots, E_n\}$, the users now remember the following values:

- For $i = 1, \dots, n-2$, $\mathcal{U}_i$ remembers $\{\hat{g} \cdot E_1, \dots, \hat{g} \cdot E_{n-1}, C_n\}$.
- $\mathcal{U}_{n-1}$ remembers $\{\hat{g} \cdot E_1, \dots, \hat{g} \cdot E_{n-1}, E_n\}$.
- $\mathcal{U}_n$ remembers $\{g_n \cdot m_1, \dots, g_n \cdot m_{n-3}, E_{n-1}, C_n, \hat{g} \cdot E_n\}$.

Evidently, if some user tries to initiate a key refreshment with these values, the operation will fail. However, $\mathcal{M}$ can bring the users into a consistent state by forging two key refresh operations herself. For this, she needs to still have control over the communications of $\mathcal{U}_{n-1}$ and $\mathcal{U}_n$, as in the original attack.

First, $\mathcal{M}$ picks two new random values $\hat{f}$ and $\hat{h} \in G$. Then, she forges a key refresh operation by sending the following values to the different users:

- To $\mathcal{U}_i$, $i = 1, \dots, n-2$, she sends

  $$\{\hat{h}\hat{g} \cdot E_1, \hat{h}\hat{g} \cdot E_2, \dots, \hat{h}\hat{g} \cdot E_{n-2}, \hat{g} \cdot E_{n-1}, \hat{f}\hat{h}\hat{g} \cdot E_n\},$$

  pretending it came from $\mathcal{U}_{n-1}$.
- To $\mathcal{U}_{n-1}$, she sends

  $$\{\hat{f}\hat{h}\hat{g} \cdot E_1, \hat{h}\hat{g} \cdot E_2, \dots, \hat{h}\hat{g} \cdot E_{n-2}, \hat{h}\hat{g} \cdot E_{n-1}, E_n\},$$

  pretending it came from $\mathcal{U}_n$.
- To $\mathcal{U}_n$, she sends

  $$\{\hat{f}\hat{h}\hat{g} \cdot E_1, \hat{h}\hat{g} \cdot E_2, \dots, \hat{h}\hat{g} \cdot E_{n-2}, C_n, \hat{h}\hat{g} \cdot E_n\},$$

  pretending it came from $\mathcal{U}_{n-1}$.

After this, the users will agree on the shared key $\hat{h}\hat{g}\cdot C_n$, which is also known to $\mathscr{M}$. As remarked above, if a user is made to believe that he received a key refreshment from $\mathscr{U}_c$, he must receive in position c the value he already held there.

Now, the values held by the users are still inconsistent, so $\mathscr{M}$ has to forge a second key refreshment:

- To $\mathscr{U}_i$, $i = 1, \dots, n-2$, she sends

$$\{\hat{f}\hat{h}\hat{g}\cdot E_1, \dots, \hat{f}\hat{h}\hat{g}\cdot E_n\},$$

  pretending it came from $\mathscr{U}_n$.

- To $\mathscr{U}_{n-1}$ and $\mathscr{U}_n$, she sends

$$\{\hat{f}\hat{h}\hat{g}\cdot E_1, \dots, \hat{f}\hat{h}\hat{g}\cdot E_n\},$$

  pretending it came from $\mathscr{U}_1$.

Now, all users and the attacker agree on the shared key $\hat{f}\hat{h}\hat{g}\cdot C_n$. Furthermore, all users remember the same consistent values for key refreshment. If in the future any user initiates a key refreshment, the attacker will lose access to the key, but the operation itself will work out without problems and without the users noticing anything wrong.

Remark 5.33. An alternative course of action for $\mathscr{M}$ is to convert the attack into a regular man-in-the-middle attack on $\mathscr{U}_n$ at the time of the first key refreshment. For this, note that given the values each user remembers, a key refreshment initiated by $\mathscr{U}_c$, $c \leq n-2$, works well for all users but $\mathscr{U}_n$. The attacker can then intercept the broadcast arriving at $\mathscr{U}_n$ and replace it with random values, except that at position n she sends $\hat{h}\cdot E_n$ for some random $\hat{h} \in G$, and at position c she sends $g_n \cdot m_c$, which she knows from step (f) of the attack. Then, $\mathscr{M}$ will have the key $\hat{g}g_c'\cdot C_n$ in common with $\mathscr{U}_i$, $i \leq n-1$, as well as $\hat{h}\cdot C_n$ with $\mathscr{U}_n$. From then on, she can run a regular man-in-the-middle attack. A similar attack can be carried out if $\mathscr{U}_n$ initiates a key refreshment, but not if $\mathscr{U}_{n-1}$ does so. In this case, the attacker can intercept and apply $\hat{g}$ to the message for $\mathscr{U}_n$ so that all users agree on a common key without noticing the previous attack.

5.7 An Active Attack on the Burmester-Desmedt Protocol

In this section, we describe a similar attack to the one in Section 5.6 for the protocol of Burmester and Desmedt [BD94; BD05]. This attack was published in [Bao+16]. We first describe the protocol in an algebraic group setting, and provide the active attack afterwards. Unlike the previous protocols, we will not translate this one to the general situation of semigroup actions.

We note that Remark 5.31 applies also to this attack: As opposed to a man-in-the-middle attack, the attacker needs to be active only during the key exchange.

Since the rekeying operation in this case is carried out by rerunning the protocol completely, an exit strategy like in Section 5.6.2 is not needed. Instead, the attacker can repeat the original attack (not necessarily on the same user) and keep listening to all communications for an unlimited time.

5.7.1 The Burmester-Desmedt Protocol

Protocol 5.34 (Burmester-Desmedt)**.** Let $\{\mathcal{U}_1, \dots, \mathcal{U}_n\}$ be a set of users that want to generate a shared key K. Let G be a group of prime order q. Let $\mathbb{Z}_q$ be the ring of integers modulo q. The users agree on a generator g of G and operate as follows:

(1) Each user $\mathcal{U}_i$, $i \in \{1, \dots, n\}$, selects a random $r_i \in \mathbb{Z}_q$ and broadcasts $z_i = g^{r_i}$.

(2) Each user $\mathcal{U}_i$, $i \in \{1, \dots, n\}$, broadcasts $X_i = (z_{i+1}/z_{i-1})^{r_i}$.

(3) Each user $\mathcal{U}_i$, $i \in \{1, \dots, n\}$, computes the key

$$K_i = (z_{i-1})^{nr_i} \cdot X_i^{n-1} \cdot X_{i+1}^{n-2} \cdots X_{n+i-2} \in G.$$

In the above, indices should be interpreted modulo n. By [BD05, Lemma 3.1], the users $\mathcal{U}_i$, $i \in \{1, \dots, n\}$ compute the same key $K = g^{r_1 r_2 + r_2 r_3 + \cdots + r_n r_1} \in G$.

5.7.2 The Attack

Under the conditions of the previous sections, let $\mathcal{U}_i$, $i \in \{1, \dots, n\}$, be a set of communicating users and let $\mathcal{M}$ be an active attacker that is able to take control of one of the users' communications, let us say $\mathcal{U}_k$. Then the attack is developed as follows.

(a) Each user $\mathcal{U}_i$, $i \in \{1, \dots, n\}$, selects a random $r_i \in \mathbb{Z}_q$ and broadcasts $z_i = g^{r_i}$ as in round (1) of the protocol.

(b) $\mathcal{M}$ stops z_k and, forging $\mathcal{U}_k$'s identity, sends to $\mathcal{U}_i$, $i \neq k$, $z'_k = g^a$, where a is such that $a - 1$ is not zero in $\mathbb{Z}_q$.

(c) At the same time, $\mathcal{M}$ stops the message z_{k+1} for $\mathcal{U}_k$ and replaces it by $z'_{k+1} = z_{k-1}^a = (g^{r_{k-1}})^a$.

(d) $\mathcal{U}_k$ starts round (2) and computes $X_k = (z'_{k+1}/z_{k-1})^{r_k} = (z_{k-1}^{r_k})^{a-1}$, which is then broadcast.

(e) $\mathcal{M}$ stops X_k and $\mathcal{U}_k$ is waiting in round (2) to receive the remaining X_i, $i \neq k$.

(f) While $\mathcal{U}_k$ is waiting in round (2), $\mathcal{M}$ finishes running the key exchange protocol with participants $\mathcal{U}_i$, $i \neq k$, using $\mathcal{M}$'s private information a. They agree on a key K.

(g) $\mathcal{M}$ computes $b = (a-1)^{-1} \mod q$ and computes $X_k^b = z_{k-1}^{r_k}$.

(h) $\mathcal{M}$ generates a list $\{h_1, \dots, h_{n-3}\}$ of elements in G and provides $\mathcal{U}_k$ the list $\{X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n\}$ given by

$$\begin{aligned} X_{k+1} &= z_{k-1}^{-r_k} h_1, \\ X_{k+j} &= h_{j-1}^{-1} h_j, \text{ for } j \in \{2, \dots, n-3\}, \\ X_{k-2} &= K\, X_k^{-(n-1)} z_{k-1}^{-2r_k} h_{n-3}^{-2} \prod_{r=1}^{n-4} h_r^{-1}, \\ X_{k-1} &= \left(\prod_{i=1, i\neq k-1}^{n} X_i\right)^{-1}, \end{aligned}$$

where indices are again taken modulo n.

After the active attack, all users $\mathcal{U}_i$, $i \in \{1, \dots, n\}$, and $\mathcal{M}$ share the same key: It is clear from step (f) that $\mathcal{M}$ and $\mathcal{U}_i$, $i \in \{1, \dots, n\}$, $i \neq k$ share the key K. A straightforward computation shows that

$$K_k = (z_{k-1})^{nr_k} \cdot X_k^{n-1} \cdot X_{k+1}^{n-2} \cdots X_{n+k-2} = K.$$

Remark 5.35. Let us note that X_{k-1} could be any arbitrary element since this is not used to compute K_k. However, in a proper execution of the protocol, it holds that $\prod_{i=1}^{n} X_i = 1$. User $\mathcal{U}_k$ could check whether this holds. In order to avoid being detected, once we have computed all X_i with $i \neq k-1$, we define $X_{k-1} = \left(\prod_{i=1, i\neq k-1}^{n} X_i\right)^{-1}$.

Bibliography

[Ahm+12] Omran Ahmadi et al. "On stable quadratic polynomials". In: *Glasgow Mathematical Journal* 54.02 (2012), pp. 359–369.

[Ahm09] Omran Ahmadi. "A note on stable quadratic polynomials over fields of characteristic two". In: *arXiv preprint arXiv:0910.4556* (2009).

[And+14] Julio Andrade et al. "Special Sets of Primes in Function Fields". In: *Integers* 14 (2014), pp. 1–4.

[AST00] Giuseppe Ateniese, Michael Steiner, and Gene Tsudik. "New multiparty authentication services and key agreement protocols". In: *IEEE journal on selected areas in communications* 18.4 (2000), pp. 628–639.

[Bao+16] Mohamed Baouch et al. "An active attack on a distributed Group Key Exchange system". In: *arXiv preprint arXiv:1603.09090* (2016).

[Bar+14] Razvan Barbulescu et al. "A heuristic quasi-polynomial algorithm for discrete logarithm in finite fields of small characteristic". In: *Advances in Cryptology–EUROCRYPT 2014*. Springer, 2014, pp. 1–16.

[BBS02] Mihir Bellare, Alexandra Boldyreva, and Jessica Staddon. "Randomness Re-use in Multi-recipient Encryption Schemeas". In: *Public Key Cryptography—PKC 2003* (2002), pp. 85–99.

[BD05] Mike Burmester and Yvo Desmedt. "A secure and scalable group key exchange system". In: *Information Processing Letters* 94.3 (2005), pp. 137–143.

[BD94] Mike Burmester and Yvo Desmedt. "A secure and efficient conference key distribution system". In: *Workshop on the Theory and Application of Cryptographic Techniques*. Springer. 1994, pp. 275–286.

[BM94a] Anjula Batra and Patrick Morton. "Algebraic dynamics of polynomial maps on the algebraic closure of a finite field, I". In: *Journal of Mathematics* 24.2 (1994).

[BM94b] Anjula Batra and Patrick Morton. "Algebraic dynamics of polynomial maps on the algebraic closure of a finite field, II". In: *Journal of Mathematics* 24.3 (1994).

[Cas86] John William Scott Cassels. *Local fields*. Vol. 3. Cambridge University Press, 1986.

[Ces81] Ernest Cesáro. "Question proposée 75". In: *Mathesis* 1 (1881), p. 184.

[Ces83] Ernest Cesáro. "Question 75 (solution)". In: *Mathesis* 3 (1883), pp. 224–225.

[Ces84] Ernest Cesáro. "Probabilité de certains faits arithméthiques". In: *Mathesis* 4 (1884), pp. 50–151.

[CL16] Joan-Josep Climent and Juan Antonio López-Ramos. "Public Key Protocols over the Ring $E_p(m)$". In: *arXiv preprint arXiv:1606.05457* (2016).

[CNT14] Joan-Josep Climent, Pedro R. Navarro, and Leandro Tortosa. "An extension of the noncommutative Bergman's ring with a large number of noninvertible elements". In: *Applicable Algebra in Engineering, Communication and Computing* 25.5 (2014), pp. 347–361.

[DH76] Whitfield Diffie and Martin Hellman. "New directions in cryptography". In: *IEEE transactions on Information Theory* 22.6 (1976), pp. 644–654.

[DM16] Edoardo Dotti and Giacomo Micheli. "Eisenstein polynomials over function fields". In: *Applicable Algebra in Engineering, Communication and Computing* 27.2 (2016), pp. 159–168.

[Dub03] Artūras Dubickas. "Polynomials irreducible by Eisenstein's criterion". In: *Applicable Algebra in Engineering, Communication and Computing* 14.2 (2003), pp. 127–132.

[Eis50] Gotthold Eisenstein. "Über die Irreducibilität und einige andere Eigenschaften der Gleichung, von welcher der Theilung der ganzen Lemniscate abhängt". In: *Journal für die reine und angewandte Mathematik* 40 (1850), pp. 185–188.

[FM15] Andrea Ferraguti and Giacomo Micheli. "On the Mertens-Cesaro Theorem for Number Fields". In: *Bulletiin of the Australian Mathematical Society* 93.2 (2015), pp. 199–210.

[FMS16] Andrea Ferraguti, Giacomo Micheli, and Reto Schnyder. "On sets of irreducible polynomials closed by composition". In: *To appear in Lecture Notes in Computer Science* (2016).

[FMS17] Andrea Ferraguti, Giacomo Micheli, and Reto Schnyder. "Irreducible compositions of degree two polynomials over finite fields have regular structure". In: *arXiv preprint arXiv:1701.06040* (2017).

[FS96] Burton Fein and Murray Schacher. "Properties of iterates and composites of polynomials". In: *Journal of the London Mathematical Society* 54.3 (1996), pp. 489–497.

[GHP99] Shuhong Gao, Jason Howell, and Daniel Panario. "Irreducible polynomials of given forms". In: *Contemporary Mathematics* 225 (1999), pp. 43–54.

[GMR81] William H. Gustafson, Marion E. Moore, and Irving Reiner. "Matrix completions over Dedekind rings". In: *Linear and Multilinear Algebra* 10.2 (1981), pp. 141–144.

[GN10] Domingo Gomez and Alejandro P. Nicolás. "An estimate on the number of stable quadratic polynomials". In: *Finite Fields and Their Applications* 16.6 (2010), pp. 401–405.

[Gni14] Oliver Wilhelm Gnilke. "The semigroup action problem in cryptography". PhD thesis. University College Dublin, 2014.

[GP97] Shuhong Gao and Daniel Panario. "Tests and constructions of irreducible polynomials over finite fields". In: *Foundations of Computational Mathematics*. Springer, 1997, pp. 346–361.

[GU82] Lucio Guerra and Emanuela Ughi. "On the distribution of Legendre symbols in Galois fields". In: *Discrete Mathematics* 42.2-3 (1982), pp. 197–208.

[GY13] Xiangqian Guo and Guangyu Yang. "The probability of rectangular unimodular matrices over $\mathbb{F}_q[x]$". In: *Linear Algebra and its Applications* 438.6 (2013), pp. 2675–2682.

[HM16] David Rodney Heath-Brown and Giacomo Micheli. "Irreducible polynomials over finite fields produced by composition of quadratics". In: *arXiv preprint arXiv:1701.05031* (2016).

[HMU01] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to automata theory, languages, and computation*. Addison Wesley, Boston, MA, 2001.

[HS13] Randell Heyman and Igor E. Shparlinski. "On the number of Eisenstein polynomials of bounded height". In: *Applicable Algebra in Engineering, Communication and Computing* 24.2 (2013), pp. 149–156.

[HS14] Randell Heyman and Igor E. Shparlinski. "On shifted Eisenstein polynomials". In: *Periodica Mathematica Hungarica* 69.2 (2014), pp. 170–181.

[JB12] Rafe Jones and Nigel Boston. "Settled polynomials over finite fields". In: *Proceedings of the American Mathematical Society* 140.6 (2012), pp. 1849–1863.

[Jon08] Rafe Jones. "The density of prime divisors in the arithmetic dynamics of quadratic polynomials". In: *Journal of the London Mathematical Society* 78.2 (2008), pp. 523–544.

[Jon12] Rafe Jones. "An iterative construction of irreducible polynomials reducible modulo every prime". In: *Journal of Algebra* 369 (2012), pp. 114–128.

[Kel55] John L. Kelley. *General topology*. New York: Van Nostrand, 1955.

[KY12] Abdel Alim Kamal and Amr M. Youssef. "Cryptanalysis of a key exchange protocol based on the endomorphisms ring End($\mathbb{Z}_p \times \mathbb{Z}_{p^2}$)". In: *Applicable Algebra in Engineering, Communication and Computing* 23.3-4 (2012), pp. 143–149.

[LLY06] Patrick P. C. Lee, John Lui, and David K. Y. Yau. "Distributed collaborative key agreement and authentication protocols for dynamic peer groups". In: *IEEE/ACM Transactions on Networking (TON)* 14.2 (2006), pp. 263–276.

[LN97] Rudolf Lidl and Harald Niederreiter. *Finite Fields*. Vol. 20. Cambridge university press, 1997.

[Lóp+15] Juan Antonio López-Ramos et al. "Group key management based on semigroup actions". In: *Journal of Algebra and Its Applications* (2015), p. 1750148.

[Lóp+16] Juan Antonio López-Ramos et al. "An application of group theory in confidential network communications". In: *Mathematical Methods in the Applied Sciences* (2016).

[Maz03] Gérard Maze. "Algebraic methods for constructing one-way trapdoor functions". PhD thesis. University of Notre Dame, 2003.

[MDM07] Johann Van Der Merwe, Dawoud Dawoud, and Stephen McDonald. "A survey on peer-to-peer key management for mobile ad hoc networks". In: *ACM computing surveys (CSUR)* 39.1 (2007), pp. 1–45.

[Mic15] Giacomo Micheli. "Cryptanalysis of a non-commutative key exchange protocol". In: *Advances in Mathematics of Communications* 9.2 (2015), pp. 247–253.

[MMR07] Gérard Maze, Chris Monico, and Joachim Rosenthal. "Public key cryptography based on semigroup actions". In: *Advances in Mathematics of Communications* 1.4 (2007), pp. 489–507.

[MP13] Gary L. Mullen and Daniel Panario. *Handbook of finite fields*. CRC Press, 2013.

[MS16a] Giacomo Micheli and Reto Schnyder. "On the density of coprime m-tuples over holomorphy rings". In: *International Journal of Number Theory* 12.03 (2016), pp. 833–839.

[MS16b] Giacomo Micheli and Reto Schnyder. "The density of shifted and affine Eisenstein polynomials". In: *Proceedings of the American Mathematical Society* 144.11 (2016), pp. 4651–4661.

[MS16c] Giacomo Micheli and Reto Schnyder. "The density of unimodular matrices over integrally closed subrings of function fields". In: *Contemporary Developments in Finite Fields and Applications* (2016), pp. 244–253.

[MW99] Ueli M. Maurer and Stefan Wolf. "The Diffie-Hellman protocol". In: *Designs, Codes, and Cryptography, Special* 20 (1999).

[Niu14] Qiuna Niu. "ECDH-based Scalable Distributed Key Management Scheme for Secure Group Communication." In: *JCP* 9.1 (2014), pp. 153–160.

[Nym72] J. E. Nymann. "On the probability that k positive integers are relatively prime". In: *Journal of Number Theory* 4.5 (1972), pp. 469–473.

[OS10] Alina Ostafe and Igor E. Shparlinski. "On the length of critical orbits of stable quadratic polynomials". In: *Proceedings of the American Mathematical Society* 138.8 (2010), pp. 2653–2656.

[Poo03] Bjorn Poonen. "Squarefree values of multivariable polynomials". In: *Duke Math. J.* 118.2 (2003), pp. 353–373.

[PS99] Bjorn Poonen and Michael Stoll. "The Cassels-Tate pairing on polarized abelian varieties". In: *Annals of Mathematics* 150.3 (1999), pp. 1109–1149.

[Rab+81] Michael O. Rabin et al. *Fingerprinting by random polynomials*. Center for Research in Computing Techn., Aiken Computation Laboratory, Univ., 1981.

[RH03] Sandro Rafaeli and David Hutchison. "A survey of key management for secure group communication". In: *ACM Computing Surveys (CSUR)* 35.3 (2003), pp. 309–329.

[Rob00] Alain M. Robert. *A course in p-adic analysis*. Vol. 198. Graduate Texts in Mathematics. Springer Science & Business Media, 2000.

[SC11] Rainer Steinwandt and Adriana Suárez Corona. "Cryptanalysis of a 2-party key establishment based on a semigroup action problem." In: *Advances in Mathematics of Communications* 5.1 (2011), pp. 87–92.

[Sch+16] Reto Schnyder et al. "An active attack on a multiparty key exchange protocol". In: *Journal of Algebra Combinatorics Discrete Structures and Applications* 3.1 (2016), pp. 31–36.

[Sch46] Theodor Schönemann. "Von denjenigen Moduln, welche Potenzen von Primzahlen sind". In: *Journal für die reine und angewandte Mathematik* 39 (1846), pp. 160–179.

[Sho90] Victor Shoup. "New algorithms for finding irreducible polynomials over finite fields". In: *Mathematics of Computation* 54.189 (1990), pp. 435–447.

[Sit10] Brian D. Sittinger. "The probability that random algebraic integers are relatively r-prime". In: *Journal of Number Theory* 130.1 (2010), pp. 164–171.

[ST07] Hiroshi Sugita and Satoshi Takanobu. "The probability of two $\mathbb{F}_q[x]$-polynomials to be coprime". In: *Probability and number theory, Advanced Studies in Pure Mathematics* 49 (2007), pp. 455–478.

[Ste+14] William A. Stein et al. *Sage Mathematics Software (Version 6.1.1)*. The Sage Development Team. 2014. URL: http://www.sagemath.org.

[Sti09] Henning Stichtenoth. *Algebraic function fields and codes*. Vol. 254. Springer, 2009.

[STW00] Michael Steiner, Gene Tsudik, and Michael Waidner. "Key agreement in dynamic peer groups". In: *IEEE Transactions on Parallel and Distributed Systems* 11.8 (2000), pp. 769–780.

[STW96] Michael Steiner, Gene Tsudik, and Michael Waidner. "Diffie-Hellman key distribution extended to group communication". In: *Proceedings of the 3rd ACM conference on Computer and communications security*. ACM. 1996, pp. 31–37.

[Syl83] James Joseph Sylvester. "Sur le nombre de fractions ordinaires inégales qu'on peut exprimer en se servant de chiffres qui n'excede pas un nombre donné". In: *C. R. Acad. Sci. Paris* 96 (1883), pp. 409–413.

[TI95] Gérald Tenenbaum and Patrick DF Ion. *Introduction to analytic and probabilistic number theory*. Vol. 46. Cambridge university press Cambridge, 1995.

[Zur03] Joachim von Zur Gathen. "Irreducible trinomials over finite fields". In: *Mathematics of Computation* 72.244 (2003), pp. 1987–2000.